Klaus Lüders
Robert O. Pohl
Editors

Pohl's Introduction to Physics

Volume 2: Electrodynamics and Optics

With videos: Demonstrations and biography of Robert W. Pohl

EXTRAS ONLINE

Springer

Pohl's Introduction to Physics

Klaus Lüders · Robert Otto Pohl
Editors

Pohl's Introduction to Physics

Volume 2: Electrodynamics and Optics

Springer

Editors

Klaus Lüders
Fachbereich Physik
Freie Universität Berlin
Berlin, Germany

Robert Otto Pohl
Department of Physics
Cornell University
Ithaca, NY, USA

Translated by
Prof. William D. Brewer, PhD,
Fachbereich Physik,
Freie Universität Berlin,
Berlin, Germany

Additional material to this book can be downloaded from
http://extras.springer.com.

ISBN 978-3-319-50267-0 ISBN 978-3-319-50269-4 (eBook)
https://doi.org/10.1007/978-3-319-50269-4

Library of Congress Control Number: 2017944730

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface to the 24th German Edition (2018) and the 2nd English Edition (2018)

Accompanying Volume 1, this second volume of “Pohl” will be published in an up-to-date format, with a modern system of numbering of the chapters, equations and figures and with exercises at the end of each chapter. Once again, we have taken the opportunity to carry out a critical reading of the whole text. Along with numerous clarifications and new formulations, we have revised many figures and comments to conform with modern notation and symbols, in order to make reviewing the material and reference to other literature as straightforward as possible.

As in Volume 1, we have carried out major revisions to the accompanying videos. In the e-book format, they will be readily accessible and can be opened directly by clicking on the links provided. This is true also of supplementary references from the Internet, and of the historical documentary “Simplicity is the Mark of Truth – the Life and Work of Robert Wichard Pohl” by the science journalist Ekkehard Sieker (see the Preface to the 22nd edition, sidenote). He has dealt extensively with the biography of Robert Wichard Pohl in Göttingen.

Once again, it is our pleasure and duty to expressly thank the *Erstes Physikalisches Institut* at the *Georg-August-Universität Göttingen* and the *Fachbereich Physik* at the *Freie Universität Berlin* for generous support. Special thanks go to Prof. K. Samwer, Prof. G. Beuermann, Mr. J. Feist and Mr. C. Mahn from Göttingen, as well as most especially Dr. J. Kirstein from the *Freie Universität Berlin*, without whose help the preparation of the videos for this new edition would not have been possible.

The first English edition of Pohl’s “Physical Principles of Electricity and Magnetism” appeared in 1930 (translated, as was his “Physical Principles of Mechanics and Acoustics” two years later, by Winifred M. Deans). It was published by Blackie & Son, Ltd., London and Glasgow. The translation was based on the second edition of Pohl’s *“Einführung in die Physik, Elektrizitätslehre”* (Julius Springer, 1929).

The present new, second English edition is based on the 24th edition of *“Pohls Einführung in die Physik”*, Vol. 2 (*Elektrizitätslehre und Optik*, Springer Spektrum, Berlin Heidelberg 2018).

Again, we gratefully acknowledge the help of Professor W.D. Brewer from the *Fachbereich Physik* of the *Freie Universität Berlin*, not only for carrying out the translation of the text with great quality and speed, but also, and this is probably even more important, for his help with the identification and clarification of unclear parts in the text and in our comments. The English-language readers will appreciate the numerous links he added for further information. We owe thanks also to E. Sieker for his translation and production of the English version of his video biography of R.W. Pohl, which is included in the links for this edition.

We also wish to thank Dr. T. Schneider and the Springer-Verlag for making this edition possible, and for their generous help in carrying out its preparation and production.

Berlin, Göttingen, Ithaca, March 2017

K. Lüders
R.O. Pohl

From the Preface to the 23rd German Edition (2009)

After the publication of Vol. 1 of R.W. POHL's *Introduction to Physics*, covering the topics of Mechanics, Acoustics and Thermodynamics in a new, revised edition in 2008, we now follow it with this Vol. 2, dealing with Electromagnetism and Optics, likewise as a newly-revised edition. We have again taken the opportunity to add supplemental information where it seemed appropriate to us. In addition to new or revised comments and a number of clarifications in the text, the novel features include in particular a series of videos showing demonstration experiments as well as a collection of exercises for the readers. The chapter on Ferromagnetism from earlier editions was again included in the section on Electromagnetism.

The additional videos for this edition were recorded under our own direction in the new physics lecture hall at the University of Göttingen, or else in cooperation with the Physics Didactics Group at the Free University in Berlin. The main part of the exercises on electromagnetism comes from the original English-language edition of 1930; however, we have again added new exercises to both the Electromagnetism and Optics sections. They deal in particular with questions which arise in the text, the figures, or the videos, and so provide additional information and an aid to understanding the concepts introduced; they are thus intended to make it easier for the reader to review and digest the material in the book.

Berlin and Göttingen, June 2009

K. Lüders
R.O. Pohl

From the Preface to the 22nd German Edition

ROBERT WICHARD POHL completed his three-volume "Introduction to Physics" in 1940 with the publication of the first edition of his "Optics". After we had edited the new, revised first volume covering the fields of Mechanics, Acoustics and Thermodynamics in 2004, we decided to combine the chapters on the fundamentals of Electromagnetism and Optics into a second volume. As with the first volume, we wanted to make an appropriate selection of the material from the many previous editions. The present Vol. 2 is based on the 20th edition of the "*Elektrizitätslehre*" and the 12th edition of "*Optik und Atomphysik*", both of which appeared in 1967. For this volume, as for Vol. 1, the *IWF Wissen und Medien* (Institute for Scientific Films) in Göttingen prepared short videos showing original demonstration experiments. They are made available with the book as a DVD.

In addition, the DVD contains the historical documentary film[1] "Simplicity is the Mark of Truth" (Original title: "*Einfachheit ist das Zeichen des Wahren*", Pohl's scientific motto; see the Comment), which was planned, researched and, together with the Düsseldorf production studio 'Kiosque', filmed by the scientific journalist Ekkehard Sieker. The film offers a detailed view of the life and work of Robert Wichard Pohl in Göttingen. It describes how R.W. Pohl, together with his famous colleagues Max Born and James Franck, made essential contributions to research and the teaching of physics in Germany in the 1920's. The Physics Institutes of the University of Göttingen in those days became one of the internationally most important centers of physics. Max Born engaged early on in research into Einstein's theory of relativity and made important contributions to the theoretical foundation of modern quantum theory. James Franck's research interests likewise lay in the field of quantum mechanics, in particular in the areas of atomic and molecular physics. R.W. Pohl was a pioneer of solid-state physics, and through his research and as a brilliant teacher, he influenced generations of physicists from all over the world. The excellent standing of physics in Göttingen was brought to an abrupt end by the political takeover of the National Socialist (Nazi) Party at the end of January, 1933. Max Born and James Franck were forced to emigrate from Germany; R.W. Pohl remained as the sole physics professor in Göttingen. He was not a publicly political person – he was a scientist, whose area of commitment was in his Institute.

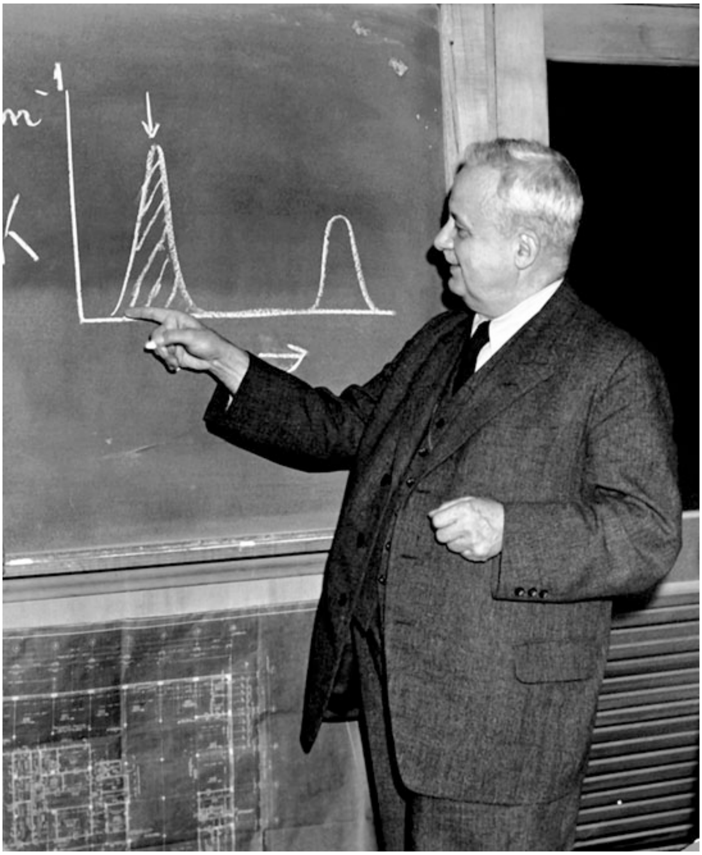
[1] **Video:**
"Simplicity is the Mark of Truth"
– the Life and Work of Robert Wichard Pohl –
http://tiny.cc/xucxny
The title of this film is the translation of Pohl's scientific motto, *"Simplex sigillum veri"*, under which he held his famous lectures over several decades. It was written on the front wall of the physics lecture hall in Göttingen, and was occasionally mistranslated by jokesters as "Sealing wax is the only truth". Pohl's successor, R. Hilsch, felt that the motto was no longer timely and had it removed during a renovation of the lecture hall some time later.
"Summer Festival 1952"
"Color Centers"
"R.W. Pohl's Farewell Lecture" (Audio only)
"Dr. h.c. for Ernest Rutherford" (Audio only)

However, in the course of his research for the film, E. Sieker came upon the little-known fact that R.W. Pohl had contacts to the civil resistance group around Carl Friedrich Goerdeler, which opposed the Nazi regime. Following the failed assassination attempt on Hitler of July 20th, 1944, his friend and contact to the Goerdeler Group, teacher and lecturer Hermann Kaiser, was sentenced to death and was executed on the 23rd of January 1945 in Berlin-Plötzensee. After the German capitulation, the British Occupying Forces made R.W. Pohl, among others, a member of the Denazification Commission for the University of Göttingen. Pohl himself regarded it as an obligation of the University to make major restitutions to the scientists who were forced to leave in 1933. The film contains many historical documents and interviews with relatives, friends and other contemporaries, giving an exceptional insight into the life of the physicist Robert Wichard Pohl.

Berlin and Göttingen, August 2005

K. Lüders
R.O. Pohl

R.W. Pohl (1884–1976)



R.W. POHL (1884–1976) discussing color centers (F-centers), elementary crystal lattice defects which were discovered at his institute and investigated there for many years. He is shown during a visit to the Ansco Research Laboratory in Binghamton, NY, in the year 1951. Details of POHL's life and work can be found on the website http://rwpohl.mpiwg-berlin.mpg.de of the Max Planck Institute for the History of Science (MPIWG). There, one can find links to other literature, scientific institutions and websites which offer information and documents on the teaching and research of the famous physicist in Göttingen. In addition, the documentary video "Simplicity is the Mark of Truth" by EKKEHARD SIEKER (Video 1) can be found on the MPIWG web site, together with all the other videos from both volumes and other audiovisual materials, available both for videostreaming or as downloads.

Contents

Part I Electricity and Magnetism

Part II Optics

List of Videos

All of the videos for this volume, and also those for Vol. 1, may be downloaded from http://extras.springer.com.

Part I Electricity and Magnetism

Part II Optics

Electricity and Magnetism

I

1 The Measurement of Electric Current and Voltage

1.1 Preliminary Remarks

Textbooks on mechanics usually begin with the concepts of *length*, *time*, and *mass*. The measurement apparatus which has been tried and tested in everyday life is described, i.e. rulers, clocks and balances, and they are immediately put to use. No one uses a sundial or a water clock for the first experiments, or even a slave who is beating time. No one begins by considering the entire historical evolution of the unit 'second'. We all make use of a modern stopwatch or digital clock. We can all use clocks, even if we are not aware of the details of their construction or of their historical development.

When the subject of thermodynamics is taken up, one generally introduces the new concept of *temperature*. The types of thermometer known to everyone in modern times are briefly discussed and they are immediately put to use in the first experiments.

In a similar way, to introduce electricity and magnetism, we make use from the very beginning of the quantities *electric current* and *voltage* which are familiar from our everyday lives. We briefly explain the instruments used to measure these quantities on an experimental basis. We then introduce the concepts of *electrical resistance, electrical energy*, and *electric power*.

Sometime in the future, this entire chapter will become dispensable. Its contents will then be just as well known from school as are the principles of clocks, balances and thermometers at present.[C1.1]

C1.1 This is perhaps already the case today, at least to a great extent. However, since this chapter gives a good introduction to the topics treated in this part of the book, we prefer not to dispense with it. We also see no difficulty when instructors choose to make use of modern, practical current and voltage sources (power supplies) and of digital measuring instruments (multimeters) for current and voltage measurements.

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_1

1.2 Electric Current

In everyday life, we speak of an *electric current* in wires and cables. We want here to start by elucidating the characteristics of electric currents. We remind the reader of two old and well-known observations:

1. Between the “north pole” and the “south pole” of a bar magnet, we can make the pattern of the magnetic field lines visible by sprinkling iron filings or powder. For example, we could put a horseshoe magnet onto a smooth surface and sprinkle the iron filings around it, tapping lightly to distribute them. We obtain the pattern shown in Fig. 1.1.

2. A magnet exerts a mechanical force on other magnets and on soft iron objects. In both cases, the field lines which we can make visible using iron filings give rise to impressive patterns. In Fig. 1.2, a horseshoe magnet is “attempting” to rotate a compass needle. In Fig. 1.3,

Figure 1.1 Magnetic field lines, represented by iron filings

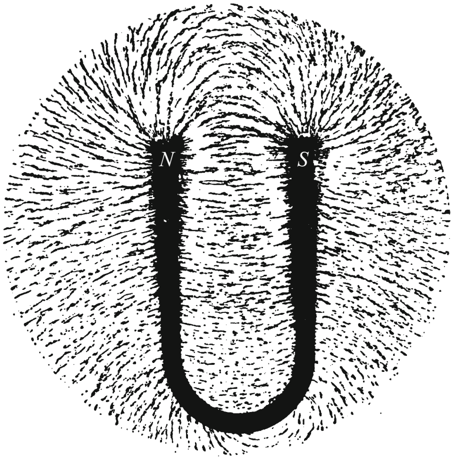
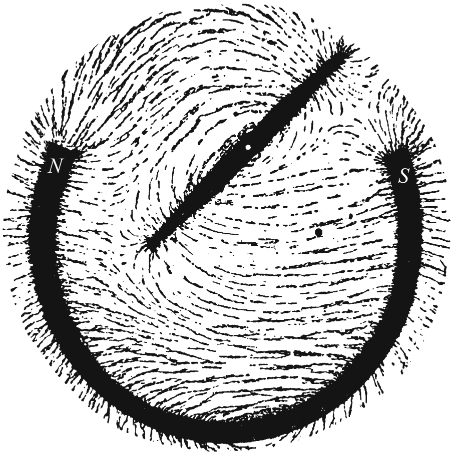


Figure 1.2 Magnetic field lines. The horseshoe magnet *SN* rotates the compass needle in a counterclockwise direction.



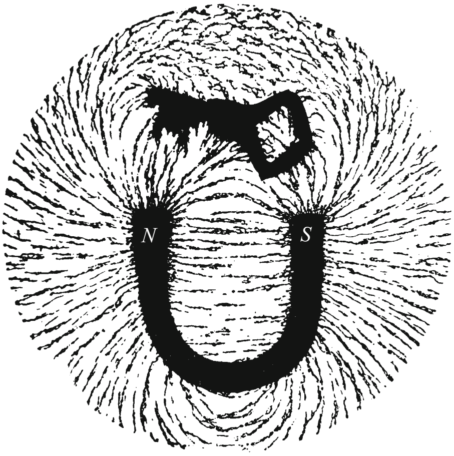


Figure 1.3 Magnetic field lines. A key is attracted to a horseshoe magnet.

a horseshoe magnet is pulling a piece of iron (a key) towards itself. We are intentionally using a somewhat primitive mode of expression here.

After these initial remarks, we now want to consider the *three defining characteristics of electric currents*:

1. *An electric current produces a magnetic field.* A wire which is carrying an electric current is surrounded by concentric circular magnetic field lines. Figure 1.4 shows these field lines using iron filings on a glass plate. The wire extended down through the page, perpendicular to the plane of the paper. It has been pulled out of the hole in the center of the picture after the field-line pattern was produced. – This magnetic field due to the current can give rise to a great variety of mechanical motions. We offer six different examples (a through f):

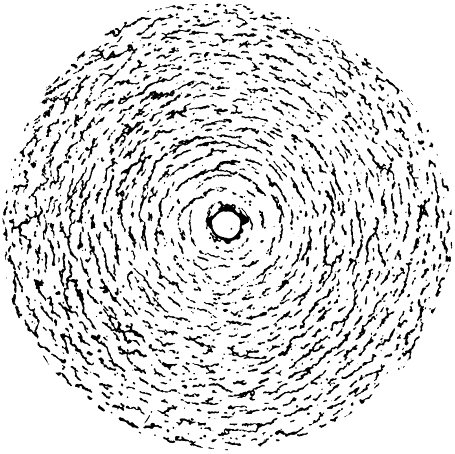


Figure 1.4 The circular magnetic field lines around a current-carrying wire

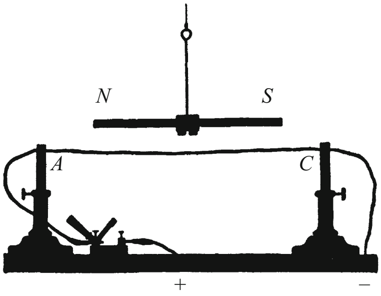


Figure 1.5 A fixed conductor (wire) *CA* and a suspended, movable bar magnet *SN*. When no current is flowing in the wire, the end of the magnet *N* points to the north. For that reason, it is called the north pole of the magnet. When the current is turned on, the north pole of the magnet is rotated out of the plane of the paper towards the observer.

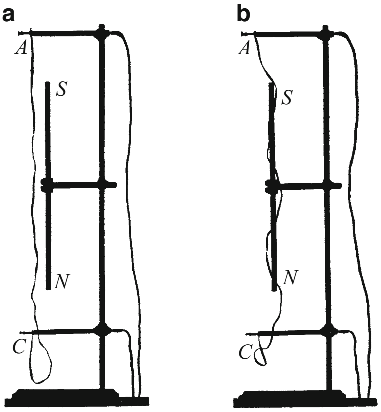


Figure 1.6 A rigidly fixed bar magnet *SN* and a movable, flexible conducting ribbon *CA* made of woven metal

a) A bar magnet (compass needle) *SN* is hung parallel above a long, straight wire *CA* (Fig. 1.5). When a current through the wire is switched on, a torque acts on the magnet and it rotates until it is perpendicular to the wire.

b) This process can be reversed. In Fig. 1.6a, the bar magnet *SN* is fixed. Beside it, a woven metal ribbon *CA* is hanging loosely; it is flexible and free to move. When a current is passed through this ribbon, it moves so that it is oriented mainly perpendicular to the magnet, i.e. it winds itself into a helix around the magnet (Fig. 1.6b).

c) We bring a straight conducting rod *CA* into the magnetic field of a horseshoe magnet *SN* (Fig. 1.7a). The rod is hung from two flexible conducting ribbons like a trapeze swing. When the current is switched on, it is deflected along one of the directions indicated by the double arrow (Fig. 1.7b).

d) We replace the straight conductor by a helically-wound conductor (coil). When the current is switched on, the coil rotates around its axis *CA* (Figs. 1.8a and b).

e) Thus far, we have considered arrangements where the magnetic field from a current-carrying conductor interacts with the magnetic field of a permanent magnet. The latter field can be produced instead

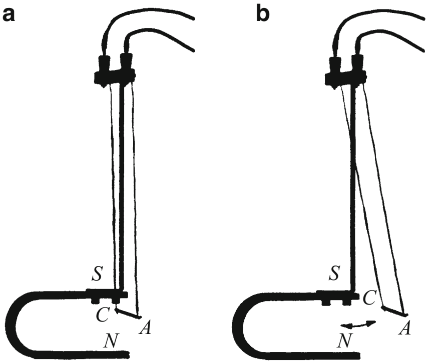


Figure 1.7 A fixed horseshoe magnet *SN* and a straight conducting rod *CA* hung like a trapeze on flexible woven-metal ribbons

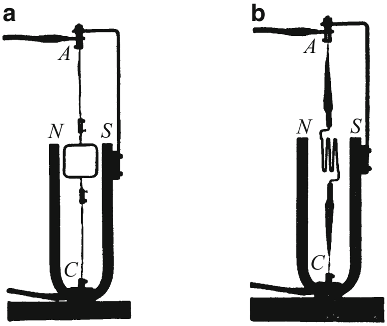


Figure 1.8 A fixed horseshoe magnet *SN* and a rotatable conducting coil *CA*. The current leads to the coil are made of flexible woven-metal ribbons. This is also the schematic of a "rotating-coil current meter" (ammeter) or "rotating-coil galvanometer".

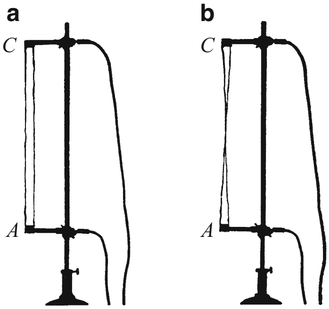


Figure 1.9 The mutual attraction of two current-carrying metal ribbons

by a second current-carrying conductor. In Figs. 1.9a and b, the current arriving at point *A* is split into two branches. At point *C*, they again flow together. The conductors along the branch *AC* are two woven-metal ribbons under a slight tension. With an electric current (Fig. 1.9b), they are attracted and are pulled together until they nearly touch[C1.2].

Figure 1.10 shows a variant of this experiment which is often used for technical applications. The two flexible ribbons are replaced by a fixed and a rotatable coil, which both carry the same electric current (Fig. 1.10a). The movable coil orients itself parallel to the fixed coil (Fig. 1.10b).

[C1.2] In liquid conductors, e.g. mercury, this magnetic force can lead to a pinch-off of the liquid column (this was demonstrated in the 21st edition of this book, p. 242).

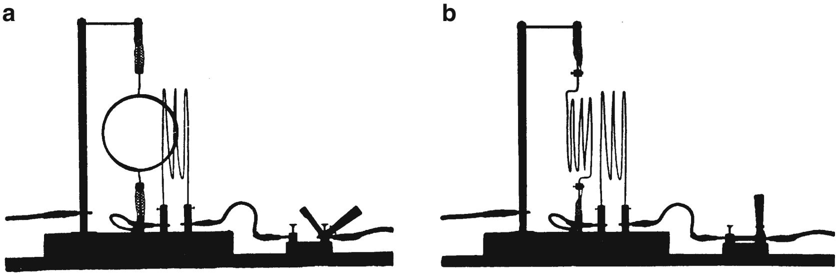


Figure 1.10 At the right is a fixed coil, at the left a coil which is free to rotate. The leads to the rotatable coil are made of woven-metal ribbons. This is a schematic of a rotating-coil instrument for measuring current or voltage, including alternating current (Sect. 10.3).

Figure 1.11 A fixed coil (solenoid) and a soft iron core *Fe* which is suspended so as to be movable

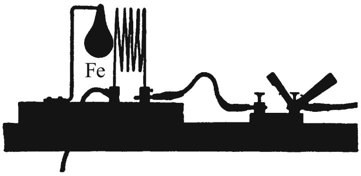
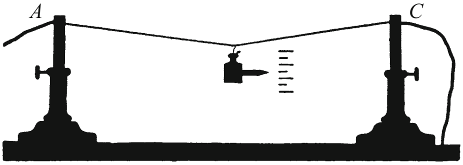


Figure 1.12 The linear expansion of a wire *CA* heated by an electric current



f) Finally, we use a piece of soft iron *Fe* (in analogy to Fig. 1.3) in Fig. 1.11. It is pulled into the magnetic field of a wire coil (solenoid). – So much for our examples of mechanical motions produced by the magnetic field of an electric current.

2. *The conductor which is carrying an electric current is heated.* It can be heated until it glows white-hot, as can be seen in any incandescent light bulb. Figure 1.12 shows a simple demonstration experiment which illustrates how a wire lengthens as a result of the heating by a current ("JOULE heating"; see Sect. 1.12). – All of the above dealt with solid conductors; we considered metal wires for the most part.

A *liquid conductor* exhibits similar effects of magnetic fields and heating. To demonstrate the magnetic field, in Fig. 1.13 a glass tube

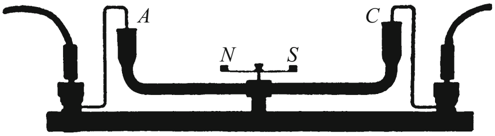


Figure 1.13 The magnetic field of a current flowing in a liquid conductor (water with a small amount of sulfuric acid added) is detected by a compass needle *SN*; paper pointers have been attached to the ends of the needle.

Figure 1.14 Electrolysis of water: The electrolytic production of hydrogen (H_2) and oxygen (O_2) when a current is passed through a dilute solution of sulfuric acid (this is a snapshot taken two seconds after switching on the current)

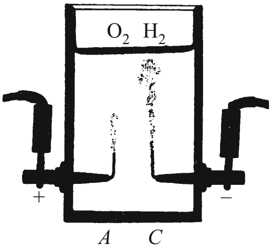
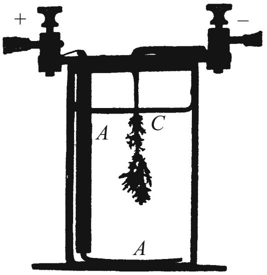


Figure 1.15 The precipitation of lead crystals at the cathode when an electric current is passed through an aqueous lead acetate solution



filled with acidified water is shown. Above it is a small compass needle. Two wires serve as current leads (*C* and *A*). – Apart from the magnetic field and the heating effect, we can observe a third effect in liquid conductors:

3. *In liquid conductors, an electric current causes chemical changes. These are termed* electrolytic. – Examples:

a) In a vessel containing acidified water, two platinum wires dipping into the water serve as *electrodes C* and *A* (Fig. 1.14). When current is flowing, small bubbles of oxygen rise from electrode *A*, and hydrogen bubbles appear at electrode *C*. *By convention, the electrode C where hydrogen is produced is called the negative pole ("cathode"). The other pole A is the positive pole ("anode")*; "*C*" for cathode, "*A*" for anode. We thus employ here an *electrolytic* definition of the difference between the negative and the positive poles in an electric circuit.

b) In a vessel containing a solution of lead acetate ("sugar of lead"), two lead wires dipped into the liquid serve as electrodes. When an electric current flows, a delicate "lead tree" made of tiny joined crystals is formed at the negative electrode *C* (Fig. 1.15). In this case, the electrolytic effect leads to the precipitation of metal out of the solution.

Finally, instead of a solid or liquid conductor, we consider a *conducting gas*. The U-shaped tube shown in Fig. 1.16 contains the noble gas neon. Again, two metal electrodes *C* and *A* serve as current leads. A small framework above the tube carries a compass needle *SN*. We attach the leads *A* and *C* to a current source; immediately, we can observe all three effects of the electric current: The magnetic needle rotates to a new position; the tube is heated; and a brilliant red-orange

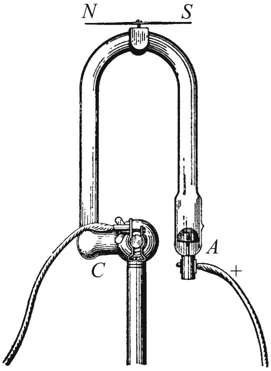


Figure 1.16 The noble gas neon as a gaseous conductor in a U-shaped glass tube. *C* and *A* denote metallic current leads, and *SN* is a compass needle.

light emitted all along the tube reveals profound changes in the gas, such as we might observe from chemical processes in a gas flame.

The *results of this paragraph*: We can characterize the effects of an electric current in a conductor by three phenomena:

$$\left.\begin{array}{l}\text{1. A magnetic field,}\\ \text{2. heating}^{1}\end{array}\right\}\text{ are produced with all conductors.}$$

3. "chemical effects" (in the broad sense) in liquid and gaseous conductors.

Or, expressed differently: We observe these three phenomena, often together, and *invent* the concept of "electric current" to summarize their cause. – This is a *qualitative* definition. It will not suffice for rigorous physics. All concepts which are employed to describe physical processes and states must be associated with *physical quantities* defined by a measurement procedure, i.e. products of a *numerical value* and a *unit*. – Here, we must be careful to keep two things separate:

1. The *definition* of a measurement procedure, and

2. the *technical setup* used for the measurements.

In the present case of electric currents, we begin with the technical setup of the measuring instruments. They can be kept quite simple: We can construct an "ammeter" which allows us to read off the strength of the current directly on a scale.

For quantitative specification, instead of simply using the word 'current', one often speaks of the *current strength* or *amperage*. This would appear to be superfluous; we do not call a measured pressure the 'strength of pressure', or a measured time the 'intensity of time', etc. But we can cite a reason for this usage in the case of an electric current: A current has a *direction*, but its strength is independent of its direction.

[1] Exception: In the case of superconducting materials, there is no heating effect.

1.3 The Technical Design of Current Meters Or Ammeters[C1.3]

For the construction of such measurement apparatus, one can use either the magnetic or the thermal effects of an electric current:

Current meters using magnetic effects (symbol shown in Fig. 1.17) are based on the arrangements illustrated in Figs. 1.5 through 1.11. Magnetic force is employed to move a pointer along a scale. The rest position of the pointer is determined by a spiral spring or some similar device. *Rotating-coil instruments* play an important role. They are based on the scheme shown in Fig. 1.8.

The magnetic field usually forms a radially-symmetric pattern; cf. Fig. 1.18, which shows two designs.

Figure 1.19a shows the coil Rc of such an ammeter with a mechanical pointer. For very sensitive instruments, a *light-beam pointer* is employed: The movable part of the instrument (rotatable coil) carries a small mirror M to reflect a light beam (e.g. a laser beam) (Fig. 1.19b). Instruments of this type are often called *mirror galvanometers*.[C1.4]

C1.3. The technical design of measurement instruments used in the laboratory today, for the most part with a digital display, is based on the principles of vacuum and solid-state physics. Such instruments will therefore be treated initially here as "black boxes", as is usual for complex apparatus in general.

C1.4. The *galvanometer*, which today may appear rather old-fashioned, is simply a particularly sensitive rotating-coil instrument, of which many are still in use. For demonstration experiments in the lecture hall, a galvanometer still offers some advantages, since it is suitable both for the measurement of weak electric currents as well as for current impulse measurements (as a ballistic galvanometer with a long response time). Furthermore, it can be used to demonstrate damped oscillations in a clear manner, including the aperiodic limit.

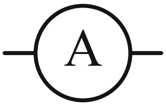


Figure 1.17 Schematic symbol for a current meter or ammeter. The same principle will later be used for instruments which are calibrated for use as voltage meters or voltmeters.

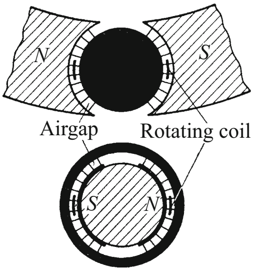


Figure 1.18 Radially symmetric magnetic fields in rotating-coil ammeters; above with the poles outwards, below with poles inwards. The magnets are shown with shading and soft iron is black. Two short circular segments mark the intersection of the rotating coil with the plane of the page.

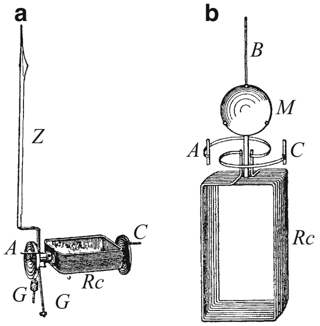


Figure 1.19 Two designs for rotating coils *Rc* in rotating-coil ammeters or "galvanometers". *C* and *A* are spiral, flexible current leads. *C* and *A* or *B* also provide the "restoring torque", i.e. they rotate the coil back to its zero position when no current is flowing. *M* is the mirror for a light-beam pointer.

1.4 The Calibration of Current Meters or Ammeters

The calibration of apparatus for the measurement of electric currents is based on the arbitrary definition of a measurement procedure and a unit for electric current. The simplest measurement procedure for comprehension and teaching was based on the *electrolytic effects* of electric currents. It makes use of the quotient

$$\frac{\text{Mass } m \text{ of precipitated material}}{\text{Time } t \text{ during which current flows}} .$$

That current which electrolytically deposits 1.1180 milligram of silver in one second was defined as the unit of current and is denoted as 1 ampere. The rather strange decimal value is due to the historical definition.

The electrolytic[2] elaboration of the unit of current called the *ampere* is especially satisfying in a conceptual sense. It states in principle: That electric current is called 'one ampere' which corresponds to the passage of a defined number of elementary electrical charges *e* through the cross-section of the current path within a given time (Sect. 3.6) (in one second about $6.24 \cdot 10^{18}$ elementary charges). The measurement of this number with the required precision by direct counting (single-electron counting) is still not possible today; therefore, one lets each single elementary electrical charge be transported by a carrier, namely a silver atom, and instead of the *number of the carriers*, one determines their total *mass* $M = 1.1180$ milligram.[C1.5] – There are naturally other procedures for the realization of the unit 'ampere'. The modern definition is based on the force between two current-carrying conductors (Sect. 8.2).[C1.6]

Many current meters, in particular rotating-coil instruments, show a pointer deflection which is directly proportional to the current; one

C1.5. Each charged silver atom (the ion Ag^+) lacks precisely one elementary charge. Silver has a molar mass of 107.87 g/mol, so that $6.24 \cdot 10^{18}$ elementary charges transport $107.87\,\text{g} \cdot 6.24 \cdot 10^{18}/6.02 \cdot 10^{23} = 1.118\,\text{mg}$ of Ag. This is an application of FARADAY's principle of equivalence, according to which the ratio of charge Q (see Sect. 2.11, Eq. 2.1, unit: ampere second, A s) to the amount of substance n deposited is given by

$$\frac{Q}{n} = z \cdot 9.65 \cdot 10^4 \frac{\text{A s}}{\text{mol}},$$

where z is the valence of the ions (= +1 for silver) (**Exercise 1.1**).

C1.6. The ampere as the unit of electric current strength belongs, along with the meter, the kilogram, the second, the kelvin, the mole and the candela, to the *base units* of the SI (Système International d'Unités), which are defined in terms of a measurement procedure. All the other units are derived from these seven base units. See the *PTB-Mitteilungen* **117**, Vol. 2 (2007); English, see http://physics.nist.gov/cuu/index.html . See also Vol. 1, Comments C2.14. and 2.15. (**Exercise 1.1**).

[2] Sometimes called the coulometric method, after the older unit for electric charge, 1 coulomb = 1 A s; now obsolete.

determines the quotient

$$D_{\mathrm{I}} = \frac{\text{Current}}{\text{Deflection}}, \qquad \text{measured in } \frac{\text{Ampere}}{\text{Scale division}}$$

to be constant and calls it the *calibration factor* of the instrument.

1.5 The Electric Voltage or Potential Difference

In everyday life, we speak of a voltage between two points, for example between the poles of a flashlight battery or the two contacts of an electric socket. – We give here the two defining characteristics of the electric voltage :

1. *A voltage can produce an electric current.* – This needs no further explanation.

2. Two bodies between which a potential difference or voltage is present are subject to mutual forces. These are often called *electrostatic forces*[C1.7].

This can be demonstrated with a force meter, e.g. a balance. In Fig. 1.20, we see a light-weight balance beam made of aluminum. It rests on a knife-edge on the metal post S. At the left end of the beam is a metal plate C, and at the right end, as counterweight, some small sliders R made of paper. Below the metal plate C is a second similar fixed metal plate A; the spacing between the two plates is a few millimeters. The plate A and the post S are each connected via a wire to the poles of a current source. When contact is made, the balance beam immediately begins to move. The voltage between the plates A and C produces an attractive force (Sect. 3.4).

So much for the qualitative properties of electric voltage or potential differences. For the purposes of physics, we must define a *measurement procedure* for voltages. Here, again, we consider the definition of the measurement procedure separately from the technical design of the instruments used for making the measurements. We will start with the latter. Both of the characteristic properties of electric voltage can be used to construct voltmeters, and we thus distinguish between instruments which carry a current, and static voltmeters ("electrometers"). We will discuss the two groups separately in Sects. 1.6 and 1.8.

[C1.7] The analogy between mechanical force and electric voltage is often emphasized by referring to the voltage as the "electromotive force" (E.M.F.). See also the footnote in Sect. 9.1

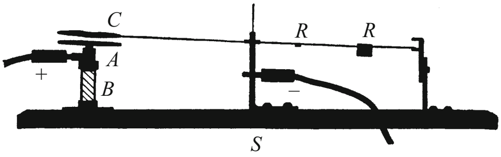


Figure 1.20 A "volt balance". B is an insulator.

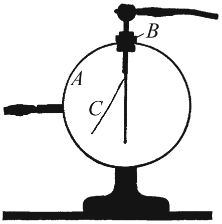


Figure 1.21 A static voltmeter with a gold-leaf pointer *C* (instruments with glass housings are practically useless; see Sect. 2.6)

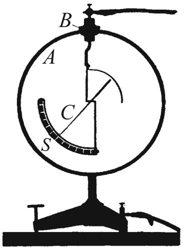


Figure 1.22 A static voltmeter (*field electrometer*) with an aluminum pointer *C* on needle bearings. It can be used for voltages between a few hundred up to about 10 000 volt.

1.6 The Technical Design of Static Voltmeters$^{C1.8}$

C1.8. Here again, we refer to the remarks in Comment C1.3. The 'electrometers' described in this section, which measure voltages with practically no current flowing during the measurement, as well as the 'rotating-coil instruments' described below (Sect. 1.8), are intuitively clear and simple in their operation, but are largely obsolete today. We will represent them in circuit diagrams by a circle containing 'V' (for voltmeter), 'E' for electrometer, i.e. a voltmeter in which no current flows during the measurement; or 'G' (for galvanometer) and 'A' (for ammeter). This will specify where relevant whether a static ('E') or a current-carrying instrument ('V', 'A', 'G') is meant.

These instruments make use of the "static" forces caused by a voltage (i.e. by the electric charges collected by the voltage; see below). They operate on the same principle as a small balance: The force caused by the voltage to be measured can be read off a scale. We mention only three of the many different designs used for such instruments:

a) The *gold-leaf electrometer* (Fig. 1.21), obsolete; strictly speaking, a gold-leaf *volt*meter. The metal housing *A* is penetrated by a metal rod, electrically isolated from the housing by an insulator *B*. The rod carries a strip of gold leaf *C* which serves as a movable pointer. The voltage to be measured is applied between the points *A* and *C*, e.g. by connecting them to a current source. The gold-leaf pointer is attracted by the wall and deflected; its angle of deflection can be read off a scale.

b) The *field electrometer* (Fig. 1.22). Its operation is similar to that of the gold-leaf electrometer, but instead of the gold leaf strip, an aluminum pointer *C* with needle bearings indicates the force which is proportional to the applied voltage.

c) The *double-fiber voltmeter* (Fig. 1.23). Here, again, a metal rod is isolated from the metal housing *A* by an insulator *B*. A loop *C* of fine platinum wires or fibers is hung from the rod; it is held under mechanical tension by the small quartz stirrup *Q* below. An electric voltage applied between *C* and *A* causes the fibers to approach the walls of the housing (or more accurately, the wire loops *A* attached to the walls). The spacing of the fibers thus increases; this increase is observed by a microscope. Figure 1.24 shows an image of its field

Figure 1.23 Model of a "double fiber electrometer". Its measurement range lies between 30 and 400 volt.

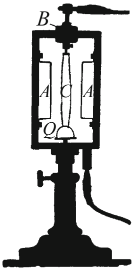
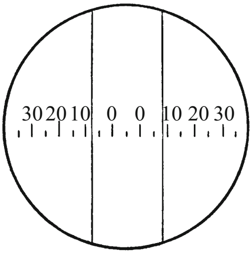


Figure 1.24 The field of view of a double-fiber voltmeter with platinum-coated quartz fibers



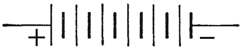
of view, with a scale. The double-fiber voltmeter is suitable for projection. Because of its rapid reaction to applied voltages, it is very convenient to use in showing demonstration experiments.

1.7 The Calibration of Voltmeters

The calibration of these instruments is based on the conventional definition of a measurement procedure and a unit for the potential difference. The *simplest* measurement procedure makes use of a *series circuit of N identical batteries* (Fig. 1.25), and asserts that the voltage between the ends of the circuit is N times larger than that of a single battery (G. S. OHM, 1827). *Within the large number of current sources, one particular battery (voltaic element) is chosen as the "normal element" and its voltage is assigned a fixed value.*[C1.9] The unit of potential difference or voltage is 1 volt (V), and all voltages are quoted in multiples of this unit.

C1.9. The volt is a derived unit; see Eq. (1.12). However, for practical reasons, it is realized by different methods, e.g. by using normal elements or, for the highest precision, by making use of a quantum-mechanical effect (the JOSEPHSON effect in superconducting junctions).

Figure 1.25 A series circuit of 6 batteries. (The positive electrode is always denoted by a longer bar symbol)



1.8 Current-Carrying Voltmeters. Electrical Resistance

Current-carrying or *galvanic* voltmeters are in principle simply recalibrated ammeters. The recalibration is made possible by the fact that in metallic conductors, a fixed relation holds between voltage and current.

In general, we define every conductor as an electrical *resistance*[3] R by the quotient

$$R = \frac{\text{Voltage } U \text{ between the ends of the conductor}}{\text{Current } I \text{ through the conductor}} \,. \tag{1.1}$$

The electrical resistance in general depends in a complex way on the current I (examples: fluorescent tubes, electric arcs, irradiated crystals, photocells). Only in special cases does one find a *constant value* of U/I, *independent of* I. One then says that Ohm*'s law* holds for that conductor:

$$U/I = R = \text{const.} \tag{1.2}$$

In words: The resistance U/I of the conductor has a constant value R, i.e. the current I through the conductor and the voltage U between its ends are strictly proportional to one another. This special case is found for metallic conductors at constant temperature[4].[C1.10]

C1.10. This type of conductor is also called an "Ohmic conductor". Suppose that it has a cross-sectional area A and a length l. Then the value of its resistance is proportional to l and inversely proportional to A. We find

$$R = \varrho \frac{l}{A} \,.$$

The proportionality factor ϱ is a property of the material and is called the *specific resistance* or *resistivity*. Its reciprocal $\sigma = 1/\varrho$ is called the *specific conductance* or *conductivity*. As an example, for copper at 20 °C: $\varrho = 1.55 \cdot 10^{-8}\,\Omega\,\text{m}$, $\sigma = 6.45 \cdot 10^{7}\,\Omega^{-1}\,\text{m}^{-1}$.

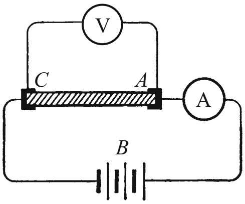
This can be demonstrated with the setup shown in Fig. 1.26. A current source B sends a current through a metallic conductor CA, e.g. a ribbon or strip of metal. The ammeter measures the current I through the conductor, while the voltmeter registers the voltage U between its ends CA. – We make use of a series of different current sources (e.g. various batteries or accumulators) and thereby vary the current I. Then we divide the corresponding values of U by I and find U/I to be constant. We thus measure the *resistance*, defined as the quotient U/I (e.g. in volt/ampere). This ratio volt/ampere is named and abbreviated by international convention as the "ohm", with the symbol Ω.

Suppose that with the setup shown in Fig. 1.26, we find for the conductor CA a value for the quotient U/I of 500 volt/ampere. Stated

[3] The word "resistance" has three different meanings within the field of electromagnetism: Firstly, it refers to the quotient of voltage and current, U/I, for any arbitrary conductor. Secondly, it refers to a *device*, e.g. a length of wire wound on a spool, as seen in Fig. 1.28; this is also called a *resistor* or *rheostat* (when it is variable). Thirdly, resistance means, as in everyday life, a force which is directed oppositely to the velocity of a moving electric charge, similar to a frictional force.

[4] The quotient mass m/volume V is defined as the *density* of an object. Under constant ambient conditions (pressure, temperature etc.), it is constant for many materials. But it is still not usual to consider the relation $m/V = \text{const} = \varrho$ as an empirically-discovered 'law' or to name it after its author.

Figure 1.26 The measurement of a resistance U/I (e.g. of the filament *CA* of an incandescent lamp) or the demonstration of the special case that OHM's law holds (the conductor *CA* could be a flat metal ribbon held at constant temperature for this demonstration)



briefly, the conductor *CA* then has a resistance $R = 500$ ohm. If two resistances (cf. footnote 3 on previous page) are connected in series, their overall resistance is equal to the sum

$$R = R_1 + R_2\,. \tag{1.3}$$

For parallel circuits, the overall resistance R is given by the equation

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} \tag{1.4}$$

(G. S. OHM). So much for the definition of electrical resistance and OHM's law.

OHM's law *now allows us to re-calibrate an ammeter to make it into a voltmeter*. – The usual ammeters contain a wire through which the current to be measured flows, for example rotating-coil instruments (Fig. 1.19). We know the value of its resistance, defined by the ratio

$$R = x\frac{\text{volt}}{\text{ampere}} = x \text{ ohm}\,;$$

x is here a numerical value. As a result, we need only to multiply the ampere calibration by the factor $R = x$ volt/ampere in order to convert the ampere calibration to a volt calibration (and the instrument into a voltmeter).

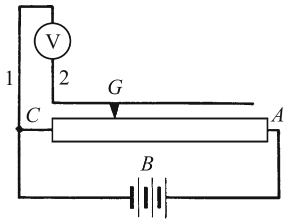
We repeat: Current-carrying voltmeters are simply re-calibrated ammeters. We therefore draw them in our circuit diagrams according to the scheme shown in Fig. 1.17, but with the letter 'V' to indicate 'voltmeter'.

The measurement instruments described in Sects. 1.3 to 1.8 operate on readily recognizable physical principles. This is a great advantage for the student of physics or electrical engineering.

1.9 Some Examples of Currents and Voltages of Varying Strengths

a) Voltages of the order of 1 V are found between the poles of dry cells or batteries for doorbells, flashlights etc.

Figure 1.27 Schematic of a voltage divider circuit (potentiometer circuit)



b) The line voltage between the contacts of wall sockets is a few hundred volts. In Europe, 220 V is usual (for alternating current, cf. Sect. 10.3).

c) Above voltages of several thousand volts (kV), there will be sparks and arcing through the air. A voltage of around 3000 V (3 kV) can produce an arc 1 mm long in air.

d) For long-distance transmission lines, voltages up to 10^3 kV are used.

e) For physics research, generators which produce voltages up to tens of MV (1 megavolt = 1 MV = 10^6 V) are available (e.g. van de Graaff or "belt generators"; see Sect. 2.9).

Many experiments require variable voltages. These can be obtained by a trick which permits selecting some fraction of a maximum voltage. One makes use of a *voltage divider* or *potentiometer* circuit (Fig. 1.27). The two poles of the current source *B* are connected to a "variable resistor" *CA*, which is typically wound with a length of fairly resistive metal wire made of an appropriate alloy and fixed on an insulating drum. Between its ends *CA* is the full output voltage of the current source. Between one end and the middle is one-half this voltage and so forth for other fractions of the length. We can thus attach a wire 1 to one end of the resistor and a second wire 2 to a metal slider *G*. Then by moving the slider along the resistor, we can produce any desired voltage from zero to the source voltage between the ends of 1 and 2. – Figure 1.28 shows a convenient setup for such a potentiometer circuit.[C1.11]

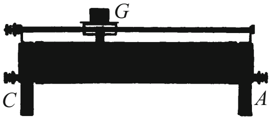
C1.11. In calculating the voltage between *C* and *G*, it must be taken into account that the current source also has an (internal) resistance which acts in series with *B* in the schematic of Fig. 1.27.

Now, we turn to some examples of current strengths.

a) Currents of the order of 1 A are passed through common lamps used for room lighting.

b) 100 A is a typical current which is employed in the motors of an electric streetcar or subway train.

Figure 1.28 The technical design of a variable resistor ('potentiometer' or 'rheostat') with a sliding contact *G*. The resistance wire is wound onto an insulating cylinder.



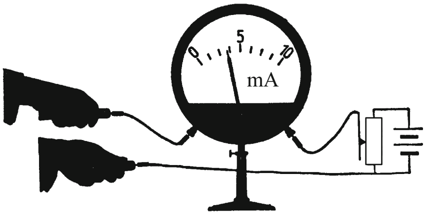
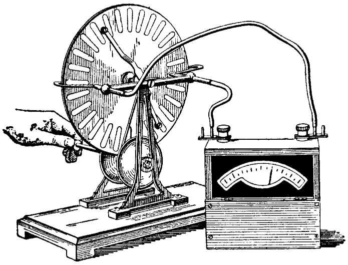


Figure 1.29 Including a test person within an electric circuit. An ammeter of the type shown in Fig. 1.8 is used. The handles contain hidden protective resistors which keep the test person safe in the event of a malfunction.

Figure 1.30 Measurement of the current delivered by a HOLTZ influence machine or "Wimshurst machine" with a rotating-coil ammeter



c) 10^{-3} A is called 1 milliampere (mA). Currents of a few mA (3 to 5 mA) can barely be perceived when they are passed through the human body. This can be demonstrated using the setup shown in Fig. 1.29. The test person is connected to the circuit via two metal handles. The applied voltage is slowly and smoothly increased using a voltage divider as described above.

d) Currents of the order of 10^{-5} A are produced by an *influence machine* or "Wimshurst machine", used in the 19th century as a high-voltage generator. We can measure this current as in Fig. 1.30 using a technical ammeter. One often encounters a strange prejudice here: An influence machine is supposed to produce "static electricity", while an ammeter measures only "galvanic" currents. In fact, there is no difference between static and galvanic electricity![C1.12]

C1.12. Here, and for all the applications mentioned later, we could replace the influence machine by an electronic high-voltage source (a "power supply"). However, just as in Fig. 1.29, it is then wise to limit the maximum current by inserting a resistance R into the circuit (as for example in **Video 3.1**).

e) 10^{-6} ampere is called 1 microampere (μA). Currents of this order of magnitude can easily be produced by the human body. In Fig. 1.31, a test person grasps two metallic handles with both hands. They are connected by wires to the ammeter (mirror galvanometer). Clasping the handles loosely, we observe no current. Tensing the finger muscles of one hand produces a current of some 10^{-6} A. If the other hand is tensed, a similar current is observed, but in the opposite direction.

f) High-quality mirror galvanometers can measure currents down to around $3 \cdot 10^{-12}$ ampere (3 pA).[C1.13]

C1.13. Using electronic instruments, today we can measure currents two orders of magnitude smaller.

This lower limit is determined by the BROWNian molecular motion of the movable parts (rotating coil etc.). With a still higher sensitivity (a lighter

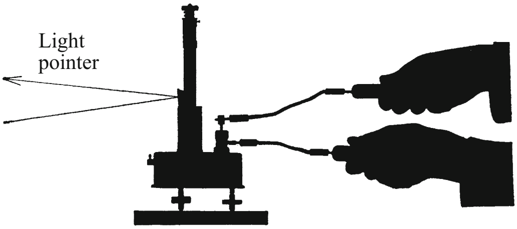


Figure 1.31 Observation of the weak electric currents produced by tensing the finger muscles. The rotating-coil galvanometer (schematic in Fig. 1.8) with a mirror and light pointer is distinguished by an especially short oscillation period ($T = 0.5$ s). (This current is due to processes in the skin and not in the muscles!)

coil or a finer suspension), the zero point of the instrument would move in just as chaotic a manner (albeit much more slowly) as the dust particles in BROWNian motion (Vol. 1, Sect. 9.1)

1.10 Current Impulses and Their Measurement

Often, in the course of physics experiments we are dealing with electric currents which are constant over time; then the pointer of an ammeter rests at a certain value on its scale and remains there, showing a constant deflection. However, in many measurements, currents flow only very briefly; for example with a time dependence like that sketched in Fig. 1.32a: The current drops within a time t from its initial value to zero. The shaded area in the figure represents the time integral over the current ($\int I\,\mathrm{d}t$). This integral has a brief and appropriate name, the *current impulse*. This term is analogous to the mechanical *impulse* ($\int \boldsymbol{F}\,\mathrm{d}t$). The simplest example of a current impulse is shown in Fig. 1.32b: A constant current I flows during

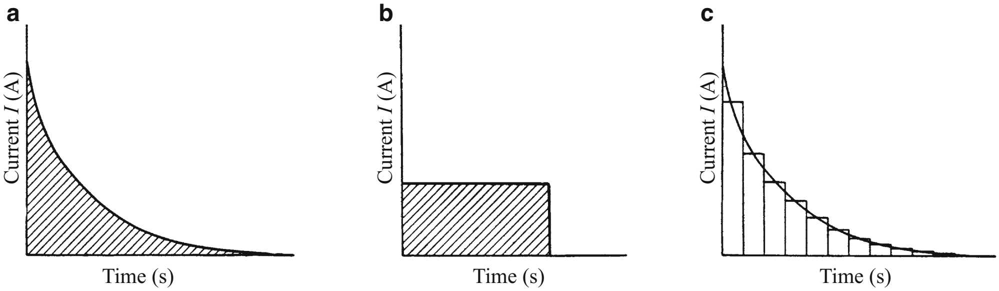


Figure 1.32 Three examples of time integrals of the current or "current impulses", measured in ampere second (A s)

a time t. The magnitude of the current impulse is then given simply by the product of the current and the time; it is thus $I \cdot t$, with the unit ampere second (A s). In a corresponding manner, by summation (Fig. 1.32c), current impulses of arbitrary time dependence can be evaluated. This is however too tedious, so that it is done only on paper.

In reality, a current impulse is a quantity which can be measured in a very convenient way. *This requires reading only a single pointer position* from an ammeter. The ammeter in this case merely has to fulfill two conditions:

1. At *constant* current, the constant deviation of its pointer must be strictly *proportional* to the current. This is to good accuracy the case for rotating-coil galvanometers (Sect. 1.3). Since one can consider a rotating-coil galvanometer to be a torsion pendulum (Vol. 1, Fig. 6.5), this proportionality means that the force produced in an ammeter of this type is proportional to the current flowing through it. The same is true of the (mechanical) impulse resulting from a current impulse.

2. The oscillation period of the pointer must be long compared to the time during which current flows in a current impulse. Then the torsion pendulum leaves its rest position with practically its maximum angular velocity, which is proportional to the impulse and thus also to the current impulse. For a pendulum with a linear restoring force law, the amplitude u_0 of its velocity is proportional to its maximum deflection x_0:

$$u_0 = \omega x_0 \,, \tag{1.5}$$

where ω is the circular frequency of the pendulum (Vol. 1, Sect.4.3).

We thus expect a constant value of the quotient

$$\frac{\text{Current impulse}}{\text{Impulse deflection}} = B_{\mathrm{I}} \,.$$

To demonstrate this, we use a current impulse with a rectangular shape (Fig. 1.32b). That is, we pass a known current I during a short but precisely measured time t through a slowly oscillating galvanometer (its period of oscillation in Fig. 1.33 is 44 s). Any suitable sort of electric time switch can be used for this purpose.

A known electric current I of suitable strength can be provided by the circuit shown in Fig. 1.33. Using a voltage divider (Fig. 1.27), for example a voltage of 1/100 V is selected. This voltage produces a current which passes through the galvanometer and through a resistance of $10^6\,\Omega$. This current I is then, according to OHM's law (Eq. 1.2), equal to $10^{-2}\,\mathrm{V}/10^6\,\Omega = 10^{-8}\,\mathrm{A}$. With this setup, we can observe the galvanometer deflection α for different values of the product It. We then repeat the measurements with different current strengths. The times t are also arbitrarily varied between a few tenths and around 2 s.

Figure 1.33 Calibration of the impulse deflections of a galvanometer with a slow response time (‘*ballistic* galvanometer’; see also Sect. 2.11) in the unit *ampere second*

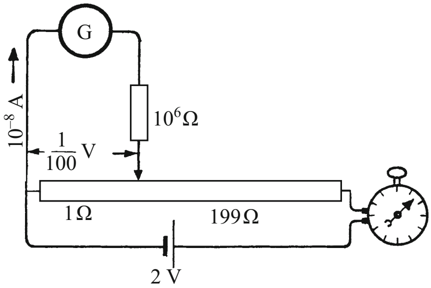
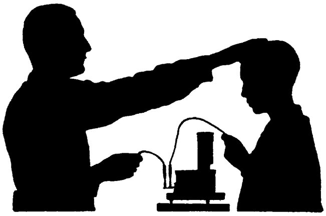


Figure 1.34 “Frictional electrification machine”. The same galvanometer is used here as in Fig. 2.36.



We then compute the ratios B_I = (Current impulse It)/(Observed deflection α). For all the measured values and in all cases, we obtain the same result; in our example, $B_\mathrm{I} = 1.2 \cdot 10^{-8}$ A s/scale division. Thus, we have demonstrated the proportionality of the observed impulse deflection to the current impulse for a current pulse of *rectangular* shape (Fig. 1.32b), and at the same time we have *calibrated* the galvanometer ballistically. This result can be readily generalized: Every arbitrary current impulse can be put together by adding rectangular pulses, as in Fig. 1.32c.

The ballistically-calibrated galvanometer can now be used to measure an unknown current impulse. To show this, we improvised a *frictional electrification machine* as shown in Fig. 1.34. Instead of sealing wax and cat’s fur, we use the hand of one of the experimenters and the hair of the other. One stroke across the hair produces a deflection of about 16 scale divisions, that is a current impulse of about $2 \cdot 10^{-7}$ A s.[C1.14]

1.11 Current and Voltage Instruments with a Short Response Time. Cathode-Ray Tubes

Great physical achievements of past decades are today already part of our general and technical knowledge, and are even topics for science fair projects and the like; some of them are also in the meantime obsolete. The cathode-ray tube (“BRAUN’s tube”) belongs to a considerable extent in this latter category.[C1.15] Developed in 1899, it

C1.14. Of course, today we have technical instruments for measuring current and voltage impulses of far greater sensitivity and precision. However, since the ballistic galvanometer contains a lot of interesting physics (e.g. it is a damped oscillator; see Vol. 1, Sect. 11.10) and is especially well suited for demonstration experiments in the lecture room, as can be seen in several of the videos for this volume, it seems opportune and reasonable to treat it in some detail at this point.

C1.15. Named after KARL FERDINAND BRAUN (1858–1918). Today, cathode ray tubes for viewscreens and television screens have been largely replaced by flat screens making use of liquid crystals or light-emitting diodes. For a discussion of “BRAUN’s tube” see e.g. F. Hars, “Hundert Jahre Braunsche Röhre”, *Phys. Bl.* **54**, 1040 (1998); English: see https://en.m.wikipedia.org/wiki/Cathode_ray_tube.

represents a current- and voltage-measuring instrument with an extremely short response time. In the oscilloscope, which is today indispensable in many laboratories, it is used to display and record rapidly-occurring electrical processes. The "pointer" is an electrically deflected electron beam which strikes a fluorescent screen and becomes visible there. Frequencies up to 10^{10} Hz can be visualized in this way.

Using deflections in two coordinate directions, one can measure two different quantities simultaneously, for example two currents, two voltages, a current and a voltage, or current *vs.* time; the time can be represented either as a linear deflection or as an angle, etc.

1.12 Measurements of Electrical Energy

Today, we can hardly imagine everyday life without electrical phenomena, with their innumerable applications. Everyone needs at least two electrical concepts in daily life, namely electric current I and electric voltage U. Both are measured as multiples of the units ampere and volt. – Making use of these two *electrical* quantities, we can also measure the electrical *energy*. A typical setup is shown on the right in Fig. 1.35. We observe the same temperature rise when the product UIt has the same value (t is the time during which the current I flows). Thus, this product is a measure of the energy E, with the unit 'volt ampere second' (V A s):

$$E = UIt\,. \tag{1.6}$$

To distinguish it from other types of energy, it is often termed JOULE heating. Applying the electrical resistance $R = U/I$, we obtain the frequently-used form

$$E = I^2 \cdot R \cdot t = \frac{U^2}{R} \cdot t\,. \tag{1.7}$$

Mechanically, we measure the energy E in terms of the product called 'work',

$$W = F\,l\,. \tag{1.8}$$

(Unit: newton meter; an experimental setup is shown at the left in Fig. 1.35. – F is the force component parallel to the path l.)

A mechanically- and an electrically-produced energy are thus equal when they increase the temperatures of two identical calorimeters (Fig. 1.35) by the same amount. That occurs in the experiment when

$$F\,l = UI\,t \tag{1.9}$$

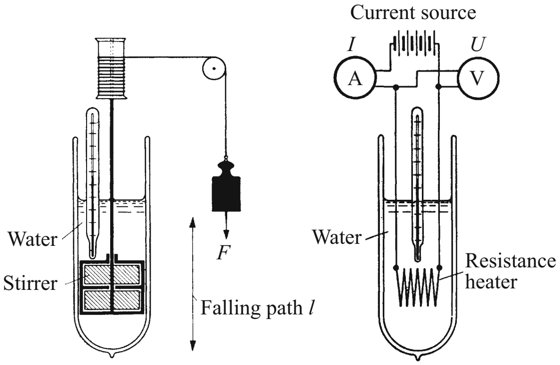


Figure 1.35 The production of identical quantities of energy, measured mechanically and electrically by means of equal temperature increases in two identical calorimeters. At the left: Mechanical energy input via a stirring apparatus; a metal block which exerts the force F falls along a path l. At the right: Input of electrical energy by heating a resistor (immersion heater).

i.e. when the product on the left with the unit N m has the *same numerical value* as the product on the right with the unit V A s. Therefore, we find

$$1\ \text{newton meter} = 1\ \text{volt ampere second}\,. \tag{1.10}$$

This equality of the mechanical and electric *energy units* is not physically necessary, but is rather the result of an expedient international agreement: The unit 'volt' has been defined in such a way that Eq. (1.10) is fulfilled. – Or, expressed differently, we dispense with defining all three of the quantities on the right in Eq. (1.9) independently of one another as *base* units. Instead, we make use of the current to measure voltage. We define the voltage U and its unit 'volt' as a derived quantity by making use of Eq. (1.9). The definition is

$$\text{Voltage } U = \frac{\text{Work } F\,l}{\text{Current } I \cdot \text{Time } t} \tag{1.11}$$

and thus

$$1\ \text{volt} = 1\,\frac{\text{newton meter}}{\text{ampere second}}\,. \tag{1.12}$$

Analogously, in the fundamental equation of mechanics, i.e. acceleration $\boldsymbol{a} = \boldsymbol{F}/m$, the force $\boldsymbol{F}$ and the mass m are not defined independently of one another as base quantities. Physicists use the mass to *measure* force; they define the force as a derived quantity with the defining equation $\boldsymbol{F} = m\boldsymbol{a}$ and the unit 1 newton $= 1\,\text{kg}\,\text{m/s}^2$.

The electrical energy unit is thus the 'watt second', so that

$$1\ \text{volt ampere second (V A s)} = 1\ \text{watt second (W s)}\,. \tag{1.13}$$

In practice, we usually employ 1 kilowatt hour (kWh) = 1 kilovolt ampere hour. This is a unit of electrical energy with an industrial price of the order of 10 Eurocents.[C1.16]

In mechanics (Vol. 1, Sect. 5.2), the concept of 'power' $\dot{W}$ was defined by the equation

$$\dot{W} = \frac{\mathrm{d}W}{\mathrm{d}t} \tag{1.14}$$

(work W, time t). Its mechanical unit is 1 newton meter/second, while its electrical unit is 1 volt ampere = 1 watt.

C1.16. The industrial price in Germany at present (2016) is about 8 Euroct./kWh; for private users, it is nearly 30 Euroct./kWh, with an increasing tendency. In the U.S., these values are 5.8 $ Ct. (industrial price) and 11.1 $ Ct. (private users), respectively, with the industrial price falling and the private-user price rising.

Exercises

1.1 For the calibration of an ammeter using the electrolytic method described in Sect. 1.4, divalent copper ions (Cu^{++}) were used. At the negative electrode, a mass increase of 5.9 g per hour was measured. The ammeter indicated a current of 4.5 A. What was the true value of the current I? (The molar mass of copper is 63.54 g/mol.) (Sect. 1.4)

1.2 Bauxite ($\mathrm{Al(OH)_3}$) is used for the electrolytic production of aluminum (the so-called flux-melt electrolysis process). The molar mass of aluminum is 26.97 g/mol. How long does it take to produce 1 metric ton of aluminum with a current of 4 000 A? (Sect. 1.4)

1.3 How could we increase the maximum indicated value of a voltmeter with the internal resistance $R_\mathrm{i} = 1\,\Omega$ from 0.15 V to 15 V? (Sect. 1.8)

1.4 The maximum indicated value of an ammeter with the internal resistance $R_\mathrm{i} = 1\,\Omega$ is to be increased from 0.05 A to 10 A. For this purpose, a resistor R_shunt is connected in parallel to its terminals (a so-called shunt resistor). Find the correct value of R_shunt. (Sect. 1.8)

1.5 In Fig. 1.26, we could replace the electrometer (static voltmeter) by an uncalibrated rotating-coil ammeter with a finite OHMic resistance of R_V. R_V is known. The voltage U is measured using the voltmeter, and the current I with the ammeter. How can we determine the resistance R of the conductor CA from these measurements? (Sect. 1.8)

1.6 A voltage U_a is applied to two OHMic resistances R_1 and R_2 which are connected in parallel, and the total current I is measured. If the resistances are connected in series, the voltage must be increased to $4.5U_a$ in order to obtain the same current. Determine R_2, if R_1 has the value $2\,\Omega$. (Sect. 1.8)

1.7 A battery with an open-circuit voltage of $U_0 = 1.5\,\text{V}$ is connected to a conductor of resistance $R = 5\,\Omega$. An ammeter with a negligible internal resistance is used to measure the current, $I = 0.25\,\text{A}$. Explain the result and calculate the voltage U_I between the poles of the battery when this current is flowing. (Sect. 1.9)

1.8 To illuminate a lecture room, 20 light bulbs are used, each operating at 220 V and 5 A (DC, from a storage battery). The battery consists of N cells connected in series, each producing a voltage of 2 V. Each cell has an internal resistance of $R_\text{Z} = 5\,\text{m}\Omega$. Calculate N. (Sect. 1.9)

1.9 An electric motor has an efficiency of 80 %, i.e. it converts 80 % of the electrical energy that is fed to it into mechanical energy. At an operating voltage of 220 V DC, it is supposed to lift a weight of mass 1.5 kg at a velocity of 2 m/s. How large is the current I that it draws? (Sect. 1.12)

1.10 In an X-ray source (see Fig. 19.10), the anticathode (anode) consists of a hollow water-cooled cylinder. In operation, it evaporates $100\,\text{cm}^3$ of cooling water per hour. The electron current in the source is 10 mA. Find the voltage U between the cathode and the anode (the cooling water is at room temperature (20 °C) when it enters, and its heat of vaporization is $L_\text{V} = 2.45 \cdot 10^6\,\text{Ws/kg}$; see Vol. 1, Fig. 14.3). The efficiency with which X-radiation is generated is very low ($<$ 1 %, see R.W. Pohl, "*Optik und Atomphysik*", 13th ed., p. 220), so that the energy transferred to the X-radiation can be neglected here. (Sect. 1.12)

1.11 a) Two wires with the resistances R_1 and R_2 are connected in series ($R_1 = 2\,\Omega$). With an applied voltage of $U = 10\,\text{V}$, the heat produced in R_2 is three times the heat produced in R_1. Find the current I which is flowing through the circuit. b) The wires are now connected in parallel and the same voltage, $U = 10\,\text{V}$, is applied. Find the currents I_1 and I_2 and the ratio of the amounts of heat produced as a function of time in the two wires, $\dot{Q}_2/\dot{Q}_1$. (Sect. 1.12)

For Sect. 3.7, see also Exercises 2.11, 2.15 and 2.16;
For Sect. 3.8, see also Exercise 2.7.

2 The Electric Field

2.1 Preliminary Remarks

The purpose of the first chapter in this book was summarized in Sect. 1.1. It was intended to provide a brief overview of the most important measuring instruments for electric current and voltage (potential difference). To this end, we introduced some of the basic concepts of electromagnetism. Now, making use of these concepts, we want to give a systematic treatment of the field of electromagnetism, essentially following its historical development. We start with the concepts of the *electric field* and *electric charge*.

2.2 Basic Observations. Different Forms of Electric Fields

Figure 2.1 shows two parallel, flat metal plates A and C. Their supports are made with insulators B, so that they are electrically isolated from each other and from the "ground". We connect these plates via two wires to a current source at a voltage of 220 V[1] and then, using two other wires, to a static voltmeter (a *double-fiber electrometer*, i.e. no current flows during the measurement). We thus have the straightforward circuit shown in Fig. 2.2, left.[C2.1] The voltmeter indicates a voltage of 220 V between the two plates. The cause of this voltage is apparently the connection of the two plates to the output poles of a current source. The experiment contradicts this assumption: The voltage remains even after the connecting wires to the current source have been removed (Fig. 2.2, right). This is extremely important!

C2.1. Such an arrangement of two electrodes insulated from one another (and usually from the ground) is called a "condenser", in this case a "parallel-plate condenser". In electronics technology, the term "capacitor" is also often used, in particular for commercially-available components with standard voltage ratings and "capacitance"; the latter is the characteristic value which quantifies their ability to store electric charge. See below for more details. In this chapter, we will use the simpler (and older) term "condenser".

Two additional experiments with the setup shown in Fig. 2.2 (right) show that there is a strong *influence of the space between the plates* on the value of the measured voltage:

1. Increasing the distance between the plates increases the voltage, while a decrease of their spacing decreases the voltage. The two fibers of the voltmeter follow the changes in the spacing of the plates

[1] Today, there are commercially-available current sources ("power supplies") with readily adjustable output voltage. In the text, we often quote a value of 220 V. This is the output voltage of a large set of storage batteries which were used in the Göttingen lecture hall for many decades; today, they have been replaced by power supplies.

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_2

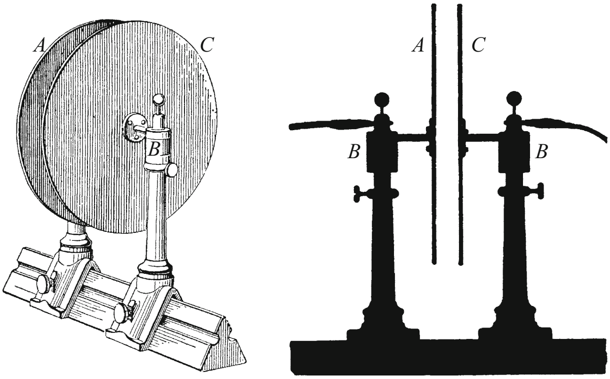


Figure 2.1 A parallel-plate condenser with insulators *B*; at the right as a shadow image. The diameter of the plates is about 22 cm

Figure 2.2 *CA*: parallel-plate condenser; at the left while it is connected to the current source, at the right after the connection is broken (here and in the following figures, the symbol for the voltmeter (circle with 'E', for *electrometer*) indicates that a *static* voltmeter (Sect. 1.6) is employed.)

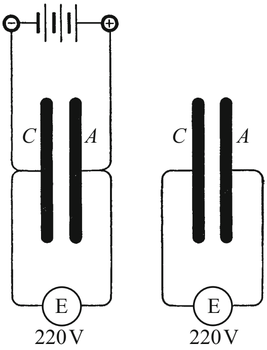
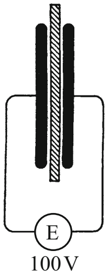


Figure 2.3 A disk made of some arbitrary material between the condenser plates (**Video 2.1**)

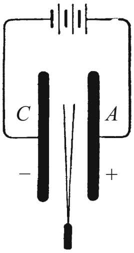


Video 2.1:
"Matter in electric fields"
http://tiny.cc/s9fgoy
This video shows how inserting various materials into the electric field of a parallel-plate condenser which has been disconnected from the current source reduces the voltage between the condenser plates; the voltage is restored when the material is withdrawn (cf. Chap. 13).

with a remarkable precision. When the spacing is returned to its original value, we again find the initial value of the voltage, in our example 220 V.

2. Without touching the plates, we slide a thick disk of some material into the space between the condenser plates (metal, plastic etc.) (Fig. 2.3). The voltage drops to a fraction of its original value. When we pull the disk out again, the original value of 220 V is restored.

Figure 2.4 Two gold-covered quartz fibers which have spread apart (the distance between the spread fibers remains small compared to the spacing of the condenser plates *A* and *C*)

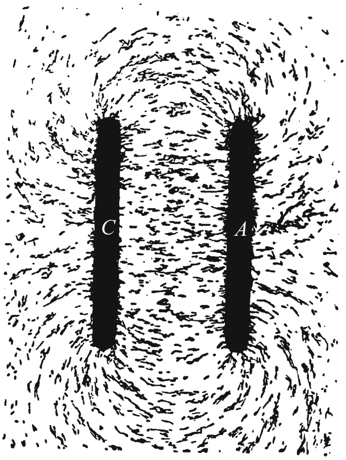


In this space between the condenser plates, unusual forces act, which otherwise do not occur; an example is shown in Fig. 2.4: Two fine metal fibers (gold-covered quartz threads) will spread apart when brought into this space between the plates (we will discuss this effect in detail in Chap. 3; see also Fig. 2.39).

We can amplify these phenomena by increasing the voltage: We replace the power supply by an influence machine or Wimshurst machine, already mentioned in Chap. 1 (Sect. 1.9); it produces several thousand volts. Then we sprinkle some small fibrous particles, for example small tufts of cotton, between the plates. The fibers stick at one end to the plates and stretch away from them. Sometimes they fly across the gap from one plate to the other, following straight-line paths in the center and curved paths at the edges of the plates. (This can be seen with particular clarity in the shadow image!)

We investigate this remarkable behavior of the fibrous particles in more detail. We try to observe it systematically in the whole of the space between the plates. To this end, we repeat the previous experiments "two dimensionally": Fig. 2.1 shows at the right a vertical section through the two plates *C* and *A*. We replace it in Fig. 2.5 by two metal foil strips glued onto a glass plate. Between them, we apply a voltage of around 3000 V. Then we scatter some sort of in-

Figure 2.5 The electric field lines in a parallel-plate condenser, made visible with gypsum crystals (this and all the following images of electric field lines have not been retouched at all)



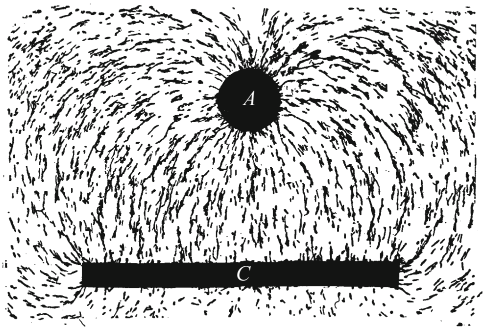


Figure 2.6 Electric field lines between a plate *C* and a sphere or a wire *A*

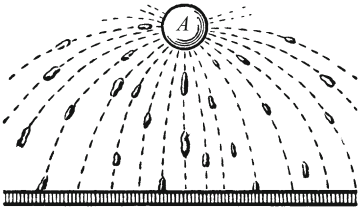


Figure 2.7 A similar image as in Fig. 2.6, from a drawing by JOHSEPH CARL WILCKE, 1777 (showing the paths of motion of gold-leaf shreds). This is the principle of "electrostatic spraying" used for painting.

sulating fibrous dust particles, e.g. powdered gypsum crystals, onto the glass plate and tap it gently. The small crystals orient themselves into peculiar lines; we see a pattern of the *electric field lines* ("lines of force"). They appear superficially to resemble the magnetic field lines which we made visible using iron filings (Figs. 1.1 through 1.4).

We can vary this experiment in a variety of ways. For example, we reshape one of the two plates into a circle or a line (wire). Then we obtain "two-dimensional" patterns as shown in Figs. 2.6 and 2.7.

On the basis of our observations thus far, we introduce *two new concepts*:

1. Two conductors between which there is an electric potential difference (voltage) are called a *condenser*.

2. The space between these two objects, in which we can detect field lines or lines of force, contains an *electric field*.

We have to derive the basic concepts of the "electrical world" from *experience*, just as we have done for the basic concepts of the "mechanical world". We have for example come to understand the phenomenon of "weight" through many experiences in daily life. Without such experiences and observations, we could not deal with mechanics. In a similar way, we will have to become familiar with the concept of the *electric field* on the basis of experience. Otherwise, we will never be able to penetrate into the world of electric phenomena. *An electric field introduces a preferred direction into a region of space; such a direction is not present in "empty" space. The field lines make this clear in a pictorial way*. In the beginning, one should approach the subject in a completely naive and unbiased manner. It is quite all right to visualize the electric field lines as visible chains of

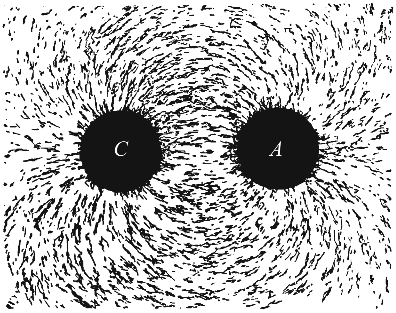


Figure 2.8 Electric field lines between two spheres or between two parallel wires

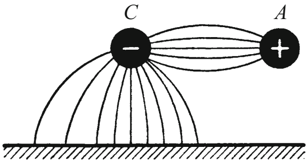


Figure 2.9 Sketch of the electric field lines between wires and the wall of the room when the positive pole of the current source is *grounded*, i.e. connected to the earth by a conducting link

dust particles. Later on, it will be no problem to distinguish between the field lines and this rough, descriptive image.

We will offer four *additional examples of condensers of different types* and show the corresponding patterns of their electric fields:

1. Two adjacent spheres or wires (Figs. 2.8 and 2.9).$^{C2.2}$

> The image in Fig. 2.8 shows how the field between the poles of our electric sockets looks. Often, one pole of a current source is permanently connected to the surface of the earth by a conducting wire. The field then looks like the sketch in Fig. 2.9. In this drawing, the positive pole *A* is assumed to be "grounded". When wires are open in the room, we sometimes see the beginnings of a "field line pattern"; one of the wires will have attracted a large amount of dust and looks like a hairy caterpillar. Along the wall adjacent to this wire is often a dusty strip; it marks the other ends of the field lines on the wall.

2. In Fig. 2.10, at the right we see a "carrier of electricity", that is, one of the plates of a condenser; it could be a metal disk *A* or a sphere. The other plate is formed by the surface of the earth, the walls of the room, the furniture and the experimenter. Figure 2.11 illustrates a dainty shape for such a carrier, a "spoon" on an insulating handle. Later, in Fig. 2.48, we will see the field of a spherical carrier of electricity.

3. An antenna and the hull of a ship (Fig. 2.12). We can see how the field lines run from the antenna to the masts and the hull.

4. Figure 2.13 shows the pattern of field lines in and around a static voltmeter. A voltmeter (or "field electrometer") of this type is simply a condenser; one of the condenser plates takes the form of a movable pointer (cf. Figs. 1.21 and 1.22).

A review of the electric fields that we have observed shows us two things:

C2.2. The concept of a *field line* has thus far simply been used to describe the experimentally observed patterns (forces on small particles) which allowed us to deduce the existence of a vector field. In graphic representations of fields, as in Fig. 2.9, one in addition makes use of the *density* of the field lines to indicate the magnitude of the field (the "field strength").

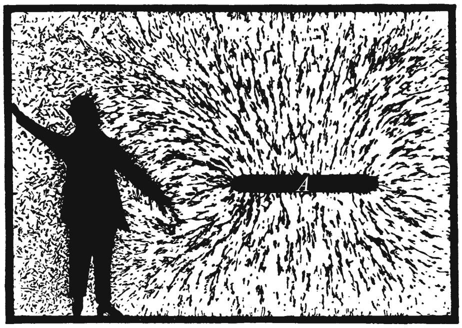


Figure 2.10 Electric field lines between a carrier of electricity (in earlier times called a "conductor") and its surroundings. J.C. WILCKE, one of the first to investigate the parallel-plate condenser, said in 1757, "The conductor namely acts as one of the plates *A*, and the observer is the other plate *C*".

Figure 2.11 A "spoon" on an insulating handle as a "carrier of electricity" or, more compactly, a "charge carrier"

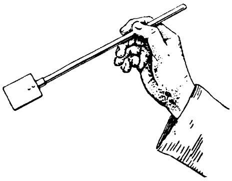


Figure 2.12 Electric field lines between the antenna and the hull of a ship (a truly "historical" picture, since the antennas for the short-wave radio signals used today are just short dipoles; see Chap. 12)

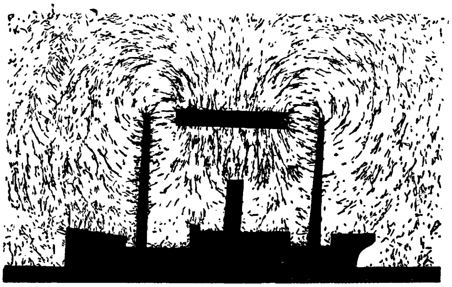
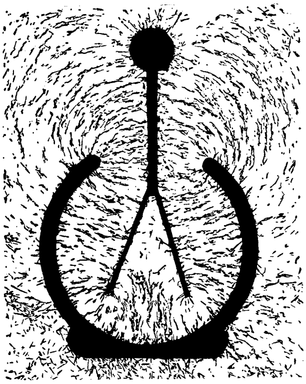


Figure 2.13 Electric field lines around a static voltmeter, or an electrometer as condenser



1. *All of the field lines always run perpendicular into the surface of the condenser plates.*

2. Among all the possible electric fields, two have a particular geometrical simplicity. *In a sufficiently flat parallel-plate condenser, the field is homogeneous* (Fig. 2.5).[C2.3] The field lines are straight, parallel, and have a constant density. A *spherical* carrier of electricity, at some distance from the other pole of the condenser, gives rise to a spherically-symmetric field (Fig. 2.48).

C2.3. The experimental proof of the homogeneity of the field can be carried out in principle by the method indicated in Figs. 2.22–2.24 using influence charges (ballistic galvanometer), whereby the two carriers of electricity must be small compared to the size of the plate (see also Sect. 2.13).

In the following, we will refer for the most part to the homogeneous field of a sufficiently flat parallel-plate condenser. The *direction* of the field will be defined as usual by the convention that *the field lines point from the positive towards the negative plate*.

2.3 Electric Fields in Vacuum

(ROBERT BOYLE, before 1694). All of the experiments described in the previous section would give the same results in high vacuum or in air. An electric field can exist in empty space. Air at atmospheric pressure has little effect on the observations of electric fields. Its influence is seen only when very precise measurements are carried out (apart from spark discharges and similar phenomena). At atmospheric pressure, a difference of 0.06 % is found for some characteristics of the field as compared to measurements in high vacuum. This result, which has been verified by numerous observations, is understandable in terms of the molecular picture of the composition of the air. Figure 2.14 reminds us of the most important facts: It shows room air with a linear magnification of around $2 \cdot 10^6$, as an *instantaneous image*. The molecules are drawn as black points. Their spherical shape is arbitrary and unimportant. Their diameter is around $3 \cdot 10^{-10}$ m. Their average spacing is about ten times larger. The volume occupied by the air molecules themselves is thus practically vanishingly small compared to the volume of empty space around them.[2]

We add to this static picture of the air a *fast exposure* with an exposure time of around 10^{-8} s (Fig. 2.15). The flight paths of three molecules are drawn in, but here only at a $6 \cdot 10^4$-fold magnification. The straight lines are the "free paths" between two collisions (about 10^{-7} m). Every kink in the paths corresponds to a collision with one of the other molecules (not shown). Their average path velocity at room temperature is about 500 m/s. 1 m^3 of room air contains around $3 \cdot 10^{25}$ molecules.

[2] In one cubic centimeter of room air, there are about $3 \cdot 10^{19}$ molecules. The diameter of each individual molecule is of the order of $3 \cdot 10^{-10}$ m. If they were strung together, they would form a chain which could be wound roughly 200 times around the earth's equator.

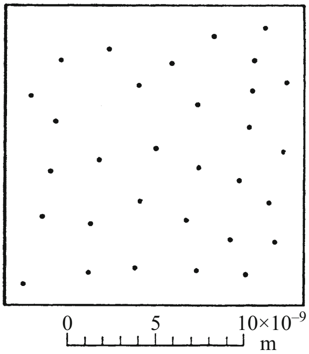


Figure 2.14 A schematic instantaneous image of room air at a $2 \cdot 10^6$-fold magnification. A cross-section of width $d = 4 \cdot 10^{-9}\,\mathrm{m} = 4\,\mathrm{nm}$ is shown.

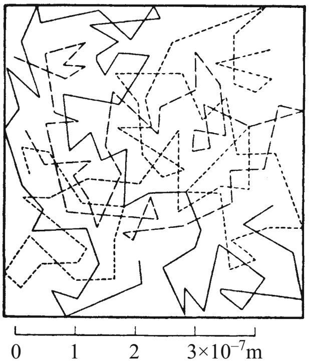


Figure 2.15 The mean free path of gas molecules in room air (magnified $6 \cdot 10^4$-fold; see also Vol. 1, Sects. 16.3 and 17.10)

2.4 The Electric Charge

We continue our experimental observations of electric fields and arrive at the following conclusion, which was already anticipated in the previous sections: *An the ends of the field lines, there is something which can be transferred or filled from one conductor to another. We will call it "electric charges"*. We will have to distinguish two types of charges (Charles F. Du Fay 1733), following a suggestion by G. C. Lichtenberg, (Göttingen 1778), who denoted them by the mathematical symbols[3] + and −. From among the many possible demonstration experiments, we give two examples:

1. In Fig. 2.16, a potential difference of 220 V has been established between the plates of a condenser by connecting them briefly to the + and − poles of the current source. Then we insert a disk-shaped carrier of electricity or charge carefully between the plates (Fig. 2.11) and move it back and forth in the direction of the double arrows. At the end of each motion, we allow the carrier disk to touch the surface of the plate briefly. Each touch reduces the voltage between

[3] They could serve equally well for distinguishing *two* different types of charge (one positive and one negative), as for an excess or a deficit of only *one* type of charge.

Figure 2.16 A charge carrier transports charges

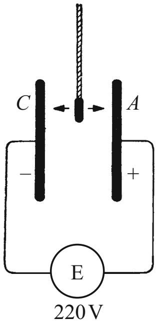
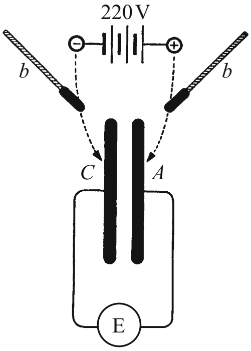


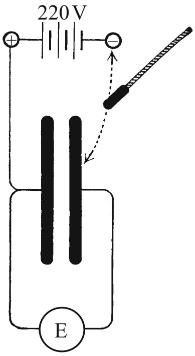
Figure 2.17 Filling electric charges from a current source (Sect. 2.2) onto the plates of a condenser



the plates. The carrier transports negative charges from the left to the right plate, and positive charges from the right to the left plate.

2. In Fig. 2.17, top, we see the + and − poles of the current source; at the bottom of the figure, there is a parallel-plate condenser connected to an electrometer, but now without a voltage between the plates. We then move two small charge carriers in the direction of the arrows along the dashed paths. We can observe that a voltage is produced between the condenser plates, and it increases with each additional motion of the carriers. Now we cross the paths, going from the − pole to A and from the + pole to C: Then the voltage decreases again;

Figure 2.18 A simplified version of the experiment in Fig. 2.17. The positive charges are transferred to the left-hand condenser plate via a wire, while the negative charges are brought to the right-hand plate by a "charge carrier" (or "spoon").



charges of the “wrong” sign are transported to the plates. (Figure 2.18 shows a simplified version of this experiment.)

Every electric charge can be decomposed into tiny portions whose size is reproducibly always the same (“elementary charges”). The first elementary charge carriers to be discovered had a negative sign and were called *electrons* (see Sect. 3.6). For clarity, we mention this fact here; we will return to it later.

2.5 Fields in Matter

We produce an electric field in a condenser in the usual manner and then bridge the gap between the plates by some object (Fig. 2.19). We repeat the experiment numerous times with different substances, e.g. in that order metal, wood, cardboard, cloth, glass, hard rubber or plastic, amber. In each case, the result is qualitatively the same: the electric field decays; the voltage between the plates drops to zero. Quantitatively, however, we find striking differences: Metals destroy the field very quickly, so that the fibers of the voltmeter collapse together in an unmeasurably short time. With wood, this process lasts several seconds, with cardboard or cloth still longer. With hard rubber or plastic, it takes quite a few minutes, and with amber, the decay of the field occurs only after some hours or days.

In this way, we can order the different materials in a series, which is called the series of decreasing electrical *conductivity*.[C1.10] The first members of the series are called good conductors, while those at its end are insulators (very poor conductors).

There is no “perfect” conductor and no “perfect” insulator. No conductor is so ideal that it destroys the field instantly. And every *insulator conducts to some extent*, i.e. it will also cause the field to decay, even if this takes a very long time.

An electric field could exist between two objects for a practically unlimited time only if they were very cold and were located in a perfect vacuum with no radiation of any kind.
The characterization of materials as ‘conductors’ or ‘insulators’ is due to STEPHAN GRAY (1729). The fact that the transition between the two kinds of material is continuous was first recognized by FRANZ ULRICH THEODOR AEPINUS (1759).

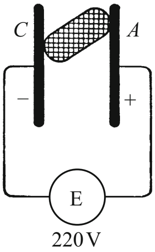


Figure 2.19 An object is bridging the gap between the two condenser plates

2.6 Charges Can Move Within Conductors

We refer directly to the preceding experiments and ask the question: How can the objects placed in an electric field destroy it? A first answer, which suffices for many purposes, can be seen by comparing Fig. 2.19 to Fig. 2.16.

In Fig. 2.16, electric charges were transported from one plate to the other using a *carrier* – the negative charges from the left to the right, and the positive charges from the right to the left. This permits the charges to combine in pairs and remain close together. Then their electric field no longer acts outside their immediate location, and it disappears in the space between the condenser plates.

In Fig. 2.19, the field disappears when the gap between the condenser plates is *bridged* by an object. This allows us to arrive readily at the following conclusion: Charges can somehow move, pulled by the force of an electric field, through material objects. The positive and the negative charges approach each other and combine as pairs. Briefly stated: *In conductors, electric charges are mobile*.

In insulators, one must then assume that there is no appreciable charge mobility within the material. This is confirmed by experiment. One can demonstrate that *electric charges stick on or in insulating materials* in numerous ways (see for example the 21st edition of POHL's *Elektrizitätslehre*, Chap. 25). We limit ourselves here to two examples:

1. We repeat the charge transfer experiment shown in Fig. 2.18, but now instead of a metal disk as charge carrier on the "spoon", we use some good insulator, e.g. plexiglas. Furthermore, in Fig. 2.20, we use a different voltage source, a small influence machine, and as a static voltmeter, we employ the field electrometer which was shown in Fig. 1.22. These two carriers of electricity behave in quite different ways. A conducting metal spoon need only touch the terminal of the voltage source at *one point*, both for taking on charge and for dispensing it. The carrier made of insulating material, however, gives rise to only very small deflections of the instrument if it is touched at just one point. In order to transfer larger quantities of charge, we have to rub the whole surface of the carrier on the pole of the current source and on the condenser plates to take up and release the charge. We have to "spread around" the charges onto the surface of the carrier in order to collect them, and to dispense them, we have to "scrape them off" again.

2. We can also make *spots* of electric charge on the surface of insulating materials. These spots can be made visible just like grease spots on a piece of cloth, by dusting them. For example, we can put an insulating plate of glass between a sheet of metal and a pointed wire. The sheet is connected to one pole of a current source with a high voltage, e.g. an influence machine. Then a small spark is allowed to

C2.4. Georg Christoph Lichtenberg (1742–1799), professor of experimental physics at the University of Göttingen (from 1770). A number of original examples of his apparatus can still be seen today in the Historical Collection at the 1st Physics Institute of the University of Göttingen.

Video 2.2: "Lichtenberg's figures"
http://tiny.cc/99fgoy
Various figures are shown which are formed when negative electric charges are either sprayed onto a plate, or are removed from it (in the latter case, an image like that in Fig. 2.21 is seen). They are compared with the images made by Lichtenberg from the Historical Collection. At the end of the video, we demonstrate how removing electrons from the interior of a plexiglas disk leads to the same kind of branched figures as on the surface. Energetic electrons were first injected to a depth of ca. 1 cm into the disk; then by grounding at the edge of the disk, they flowed out along branched paths, similar to a lightning stroke, leaving an optical change in the plexiglas so that their paths become visible.

C2.5. Influence plays an important role in producing high voltages, as we have already mentioned several times in relation to the influence machine (or Wimshurst machine). Its principles of operation were described in detail in previous editions of this book. See e.g. www.powerlabs.org/wimshurst.htm.

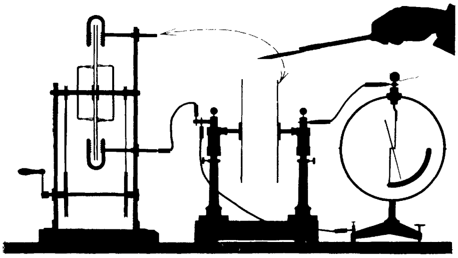


Figure 2.20 Transferring charges using charge carriers made of different materials (at the left is an influence machine which was especially designed for shadow projection)

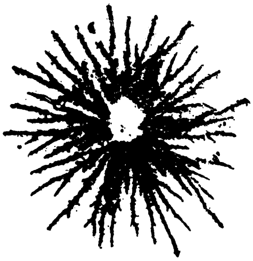


Figure 2.21 An electric spot. This kind of "Lichtenberg figure"C2.4 can also be readily produced on photographic emulsions, which then, instead of being dusted with a powder, are developed in the usual way (**Video 2.2**)

jump from the other pole of the source to the pointed wire. Initially, the charges on the surface of the glass plate are invisible; but they produce an electric field which extends out into the space above the plate. If we dust a fine powder, for example flowers of sulfur, onto the glass surface, we can see that the ends of the field lines are marked by the powder which sticks there, just as with an electric wire above a white room wall (cf. Fig. 2.9). Figure 2.21 shows an image of such a "Lichtenberg figure" (Göttingen, 1777)[4].

2.7 Influence and Its Explanation

(Johann Carl Wilcke, 1757). In our experiments on field decay up to now, we have bridged the gap between the two condenser plates using a (more or less good) conductor. Continuing these experiments, we now put a *short piece* of conducting material into the electric field, so that it does not touch the condenser plates. We can then observe the phenomenon of *influence*.C2.5 Influence will prove later to be our most important tool for detecting electric fields (induction coils, radio antennas etc.). At this point, we consider the result of our experiments which we describe in advance: *A conductor always contains positive and negative charges, but in its usual "uncharged" state, there are exactly the same amounts of the two.* The overall

[4] Electric spots can also be made by spraying charges onto a thin layer of insulating material whose optical index of refraction is >2 in the visible region. The layers become conducting on irradiation (as described in detail in the 21st edition, Chap. 25); as a result, the charges flow away where the layers have been irradiated: Dust particles stick only to the spots which were not irradiated. This is the principle of xerography, used in copying machines.

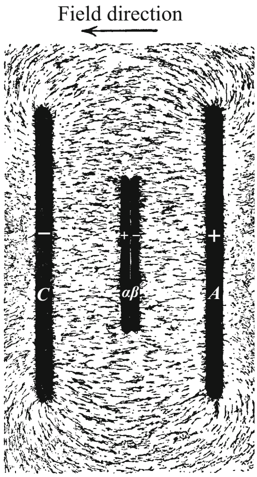


Figure 2.22 Influence. Two flat disk-shaped carriers of electricity α and β are touching each other within an electric field.

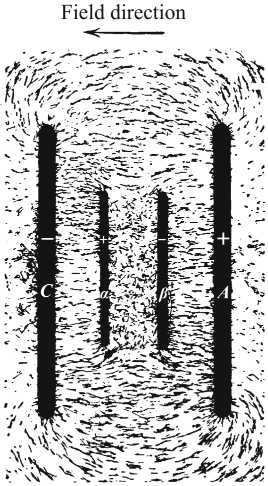


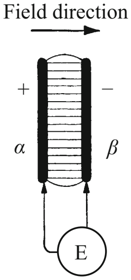
Figure 2.23 The two carriers α and β have been separated, still in the field

"charge" of an object refers simply to an excess of charges of one or the other sign.

To *demonstrate influence*, we make use of the homogeneous electric field of a sufficiently large and flat parallel-plate condenser *AC* (Fig. 2.22), and show the individual steps in the experiment in terms of two-dimensional field-line patterns. We use a metal plate as the conducting object. It is made in the form of two disks with insulating handles; the disks touch each other at several points when they are held together. The flat surfaces of the disks are oriented perpendicular to the field lines. We can then make the following observations:

1. We separate the two disks within the field and observe that the space between them is *field free*; fibrous powder shows no tendency to stick to them or to orient itself (Fig. 2.23). Interpretation: Field

Figure 2.24 Explanation of influence. The carriers of electricity which have been removed from the field region prove to be electrically charged



decay means that charges, pulled by the forces in an electric field, move within the conductors until no field remains between α and β. Where do these charges come from? The inevitable conclusion is that they were already present in the conducting disks, however as pairs (+ and −), tightly joined and thus undetectable.

2. We keep the two disks separated and remove them from the field region, as seen in Fig. 2.24; they are then connected to a static voltmeter (Figs. 1.23 and 1.24). The voltmeter indicates a voltage and thus an electric field; the two disks each carry an electric charge, with opposite signs. Interpretation: As a result of field decay within a conductor, the field lines in Figs. 2.22 and 2.23 have to end on the surfaces of the disks. The right-hand disk in these images acquired a negative charge, and the left-hand disk a positive charge.

3. How can we show that the field-line patterns in Figs. 2.24 and 2.23 are compatible with each other? Answer: The direction of the field in Fig. 2.24 is opposite to the direction of the original field in the condenser AC. The fields thus mutually cancel each other in Fig. 2.23.

The homogeneous electric field was intended only to simplify and clarify our experiments on influence behavior. In general, electric fields are inhomogeneous and the shape of objects placed in them is arbitrary. Then the field lines are not only interrupted, but also distorted, for example as in Fig. 2.25. There are always "influence charges" which collect at the points where the field lines are inter-

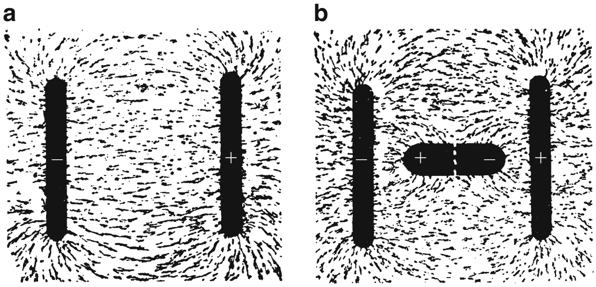


Figure 2.25 Example of influence with distortion of the electric field

rupted. These can also be detected individually in every case; one need only break apart the conductor at a suitable junction while it is in the field. This is suggested in Fig. 2.25b by the dotted line.

2.8 The Free Charges on the Surface of a Conductor

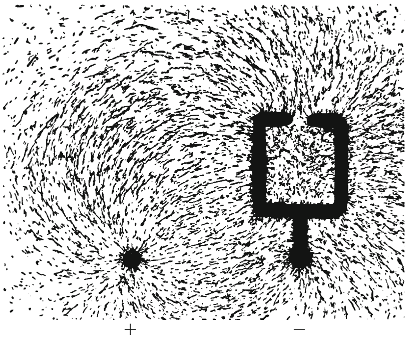
Now, continuing our experiments, we again bring a conducting object into an electric field. Previously, we bridged the gap between the two condenser plates with the object; the field decayed immediately, and we concluded that the charges within conductors can move freely. The second time, the object was located separately within the field region, not touching the plates, and we observed that the charges can be pulled apart by influence. Now, as a third case, we allow the object to touch only one of the electrodes which produce and delimit the field. We ask: How do the mobile charges distribute themselves within the conductor? We will find that the answer is that they *move to the outer surface of the conductor and remain there*.

We deduce this from a two-dimensional model experiment with field lines made visible as dust patterns. In Fig. 2.26, the two black circles indicate the poles of the current source. The field between them originally looked like the field in Fig. 2.8. Now, however, we have attached a conductor in the form of a hollow sheet-metal cup to the negative pole; it has a small hole at its top. We see that all the field lines end at the outer surface of the cup. In its interior there are no field lines, and thus also no field-line ends or free charges.

This model experiment of course requires additional confirmation through further experiments. We describe three of these:

1. Fig. 2.27 corresponds to our model experiment; but in addition, the positive pole of the current source is connected to the housing of our static voltmeter. The voltmeter acts as a condenser (Fig. 2.13), so that it can accept and store electric charge. The positive charges pass

Figure 2.26 Pattern of field lines between a sphere and a "FARADAY cup" with a small opening



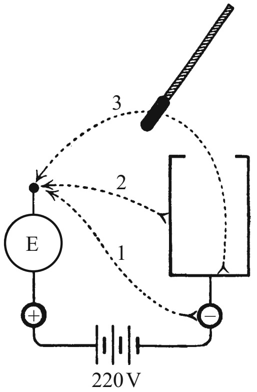


Figure 2.27 On the bottom inner surface of a conducting box which is nearly closed on all sides, or of a conducting cup, there are no free charges (BENJAMIN FRANKLIN, 1755)

through the connecting wire, while the negative charges are transported by a small "charge transport spoon" (Fig. 2.11). We first move the spoon along path 1 and observe a deflection of the voltmeter. The same behavior is found on path 2. In contrast, the spoon transfers no charge at all along path 3. This result is astonishing; the cup is connected via a conducting wire to the current source, but nevertheless, we can not collect even a small charge from its inner surface. *On the inner surface of a conducting cup, there are no free electric charges.*

A practical application: Often, one wishes to shield a certain space from electric fields. The phenomenon of influence, explained in Fig. 2.26, suggests a general method for obtaining such a field-free space: The region to be shielded need only be surrounded by a conducting shell, closed on all sides. Then the field indeed moves influence charges to the outer surface of the shell; but its inner surfaces remain completely free of charges, and its inner volume is therefore field free. The shell or shield need not even be completely closed; a housing made of wire mesh whose openings are not too large will suffice. This kind of shield is called a "FARADAY cage"). This is illustrated by the arrangement shown in Fig. 2.28.

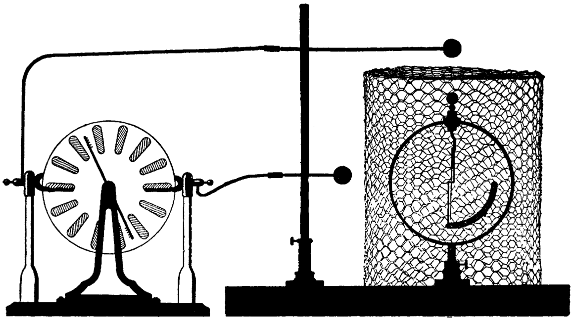


Figure 2.28 Shielding against electric fields by a metal mesh (J. S. WAITZ, 1745. The voltmeter is similar to the one in Fig. 1.22)

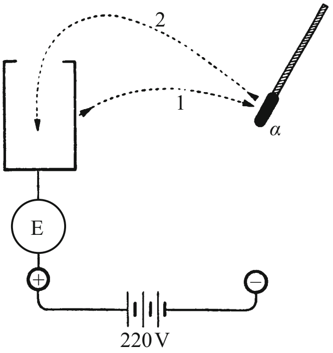


Figure 2.29 Removing and returning charges with a "charge spoon"

Without the "cage", the static voltmeter would exhibit a large deflection. Within the cage, it indicates no voltage at all. The field cannot penetrate into the interior of the cage. We can increase the voltage of the influence machine and allow loud sparks to jump between the balls and the cage, but the interior will remain free of fields and no sparks will be seen there. For sparking to occur, an electric field must be present. Protection by Faraday cages plays a significant role in the laboratory and in technical applications.[C2.6]

2. In a second experiment, we put a metal cup onto our static voltmeter (Fig. 2.29). The housing of the voltmeter remains connected to the positive pole of the current source, while the cup can be connected briefly to the negative pole to charge it. The voltmeter then indicates a voltage of 220 V. We touch the outer surface of the cup with our "charge spoon" and move the spoon about 1 m away to *a*. The voltmeter indicates a reduced voltage, i.e. some of the negative charges stored on the cup and the voltmeter have been transferred by the spoon to *a*. Then we bring the spoon along path 2 to the interior surface of the cup and fill all the negative charges back. The voltmeter again indicates 220 V. The spoon can hold no charges as long as it is connected to the interior wall of the cup, so that it is uncharged when we take it out again.

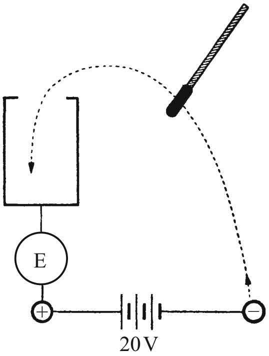


Figure 2.30 The production of a high voltage between the cup and the housing of the voltmeter (the experimenter must be aware of the meaning of Eqns. (2.3) and (2.20)!)

C2.6. Earlier editions of this book contained the following remark at this point, and the author often repeated it in private:

"Technical applications make use of Faraday cages as protection against lightning strikes. They are used for example around depots for explosives or flammable substances, in the form of a wire mesh with relatively large openings. However, another safety rule is that no ungrounded water lines from hydrants are to be passed into the enclosure; otherwise, a lightning bolt could jump from the wire cage onto the water line and thus enter the building, causing a calamity. In practice, much sad experience has been accumulated with such arrangements!"

A closed space with an insulated conductor which enters it is a *condenser*. This should be kept in mind, in particular when we are dealing with the rapidly oscillating electric fields associated with electric vibrations.

3. Finally, we carry out a third experiment with the same setup, but using a current source with only a low voltage, e.g. 20 V as in Fig. 2.30. We move the spoon back and forth between the negative pole and the inner surface of the cup. We can thereby increase the voltage indicated by the voltmeter as much as we wish, for example up to around 400 V, the maximum allowed voltage for the double-fiber voltmeter. Explanation: *In the interior of the cup, all of the charge on the spoon is deposited each time*. There is no opposing electric field there to limit the amount of charge delivered. This trick is used technically in the construction of high-voltage generators, as described in the following section. It is also important for the operation of influence machines.

2.9 Current Sources Delivering High Voltages

For voltages of up to around 5 million volts (5 MV) in air, *van de Graaff generators* or belt generators are used; they are constructed as shown in the schematic in Fig. 2.31. They are used e.g. for low-energy nuclear physics research (production of artificial isotopes). The field is generated between two large, spherical electrodes A and C. A is connected to the $+$ pole of a small battery. The other pole of the battery "sprays" negative charges via a conducting brush 1 onto a moving carrier of electric charge; this is an endless belt which is driven by a small electric motor. The charge on the belt is carried into the interior of the hollow sphere C, where it is removed by brush 2; all of it flows out onto the outer surface of the sphere. Belt generators of this type have been constructed with spherical electrodes of up to several meters in diameter, so that the experimenter can sit safely in the field-free interior while making observations.

One can leave off the small battery and use *frictional electricity* ("static electricity") to charge the belt, employing the friction between brush 1

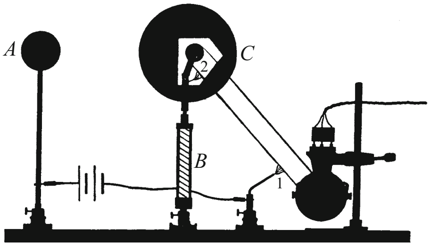


Figure 2.31 A belt generator for high voltages without corona discharge losses. The interior of the sphere C is visible through two windows. (B is an insulator; at the lower right is an electric motor.) Sparks up to 30 cm long can be produced, corresponding to a voltage of $\approx 10^6$ V.

and the belt. This is in fact simply a newer version of the old frictional electrification machine (OTTO VON GUERICKE, 1672). The rotating carrier of electricity is no longer a sphere or drum, but rather an endless belt (WALKIERS DE ST. AMAND, 1784, R.J. VAN DE GRAAFF, 1933). Carriers of electric charge in the form of belts can be made larger than disks, spheres or drums, and thus permit greater charge separation distances and higher voltages.[C2.7]

C2.7. Modern van de Graaff generators for research or isotope production have their high-voltage parts enclosed in a pressure tank containing an insulating gas (often SF_6) to prevent sparking; they can produce up to 25 MV. Even higher voltages can be obtained by operating two or more in tandem (i.e. in series). Replacing the rubber or textile belt by a chain with alternating insulating and conducting links (the "pelletron") permits a higher operating velocity and thus still higher voltages and currents. See e.g. https://www.aps.org/publications/apsnews/201102/physicshistory.cfm.

2.10 Currents During Field Decay

Based on our observations, we have attributed the decay of electric fields to a movement of charges within matter (conductors). We will now try to gain some more detailed knowledge of this charge movement, and we find that: *During the decay of the field, an electric current flows through the conductor*. We can observe this current using a technical ammeter, e.g. a mirror galvanometer with a rapid response time. We use a large condenser as shown in Fig. 2.32, consisting of 100 pairs of plates (with an area of all together about 8 m^2 and a spacing of 2 mm between the plates; cf. Fig. 2.52). We connect a voltage of 220 V to this condenser, as usual. Then its field is caused to decay by a piece of conducting wire. The galvanometer is connected in this wire in series with a piece of wood. The wood acts as a poor conductor and slows the decay of the field, so that it takes about 10 s. During this time, the galvanometer deflection indicates that a current is flowing. The time evolution of the current can be recorded with the aid of a stop watch and is shown in Fig. 2.33. The

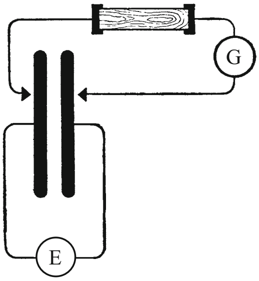


Figure 2.32 The slow decay of an electric field through a poorly-conducting wooden link (the static calibration factor of the galvanometer is $B_\text{I} \approx 2 \cdot 10^{-7}$ A/scale division)

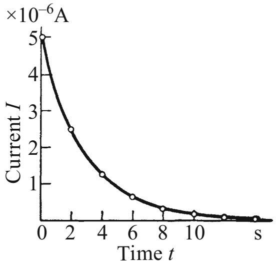


Figure 2.33 The current which flows during the decay of the field (recorded using a rapidly-responding galvanometer as in Fig. 1.31)

Figure 2.34 During field decay through a conducting wire 1, a small incandescent lamp included in the circuit lights up

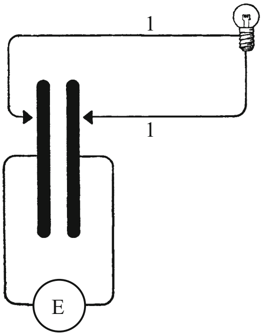
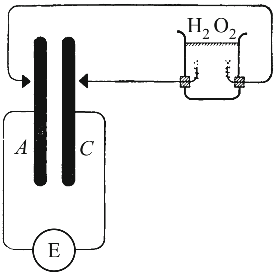


Figure 2.35 During field decay through a conducting wire, an electrolytic cell included in the circuit exhibits electrolysis (electrode area < 1 mm^2)



quantitative treatment of this curve will be given in Sect. 2.16. Of course, the brief current flow during the decay of the field could also be detected through the heat it generates or its electrolytic effects. We illustrate both using the schematics in Figs. 2.34 and 2.35.

2.11 Measurements of Electric Charge from Current Impulses. The Relation Between Charge and Current

In our experiments with electric fields, we made use of field decay (discharging of condensers) to gain some particular insights; it led us to several important phenomena: First to influence, then to the location of mobile charges on the outer surfaces of conductors, and finally to the currents which flow through a conductor during discharge. This latter phenomenon now brings us closer to an important goal, the quantitative measurement of electric charges in electrical units.

We refer to Fig. 2.33, an arbitrary example of the time dependence of the current during a field decay (condenser discharge) process. The area enclosed under the curve is the time integral of the current, that is the *current impulse* (see Sect. 1.10). We now measure this current impulse on discharging the small condenser that we have often

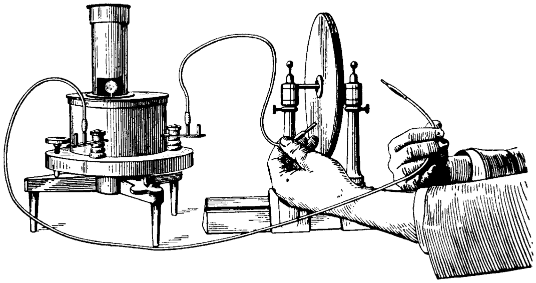


Figure 2.36 The technical arrangement used for the experiment in Fig. 2.32; at left, a mirror galvanometer. The window at the bottom of the column allows us to see the mirror which reflects the light pointer onto the scale. The oscillation period T of this galvanometer is long, around 44 s (*ballistic* galvanometer), and the spacing of the condenser plates is 4 mm.

used already, together with the *ballistic* galvanometer (with a long response time) which we calibrated in Sect. 1.10 (Fig. 2.36).

We carry out the experiment several times with variations. In all cases, the condenser plates are set initially to the same spacing, ca. 4 mm, and a field is produced (charging the condenser) by applying a voltage of 220 V (static voltmeter!). Then the experiments are as follows:

1. The connection used to discharge the condenser and destroy the field contains only the rotating-coil galvanometer with its low-resistance coil. The field decays in an unmeasurably short time.

2. A poorly-conducting (resistive) object is placed in the circuit, in series with the galvanometer, for example a piece of wood (compare Fig. 2.32). The decay of the field now requires several seconds.

3. First, the spacing of the condenser plates is increased, increasing the voltage correspondingly. Then the field is caused to decay (condenser discharge) as before, either rapidly or slowly by using the wooden “resistor”.

4. and 5. Now, we carry out two experiments on the *production* of the field. We return the condenser plates to their original spacing (4 mm), but this time, we include the galvanometer in one of the two conducting wires used to charge the condenser and thus produce the field (Fig. 2.37). In the fourth experiment, we generate the field very rapidly; in the fifth experiment, we generate it by using a resistive conductor much more slowly, over several seconds.

In *all five cases, we observe a current impulse of the same magnitude* (in this example, about 10^{-8} A s; see Eq. (2.1)). We changed the time dependence of the field, the magnitude of the voltage, we observed both the decay and the production of the field. What remained unchanged? Only the electric charges that were placed on or taken off the condenser plates, negative charges on one plate and positive on

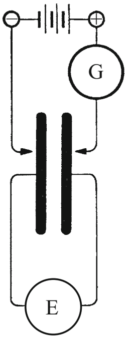


Figure 2.37 The current impulse during the production of the field

the other. From this we may deduce that the current impulse $\int I\,\mathrm{d}t$ associated with the decay or production of the field is a measure of the electric charges Q associated with the field. *We can measure electric charges Q by determining a current impulse.*

We define the charge Q through the equation

$$Q = \int I\,\mathrm{d}t \tag{2.1}$$

(Unit: 1 ampere second (A s), also called 1 coulomb (C) (obsolete)).

As a first example, we measure the *charge on a small "carrier"* (a spoon with an insulating handle), as shown in Fig. 2.38. We place a negative charge on it by touching it briefly to the negative pole of the current source. Before doing that, we had already connected the left terminal of the galvanometer (which is calibrated in ampere second) to the positive pole of the current source. Now we move the spoon along some arbitrary path to the right terminal of the galvanometer and observe a current impulse of $6 \cdot 10^{-10}$ A s. The spoon thus contained a negative charge of this magnitude.

For a quantitative investigation of the mechanisms of electrical conduction with the aid of similar experiments in gases, see for example the 21st edition of this book, Chap. 15.

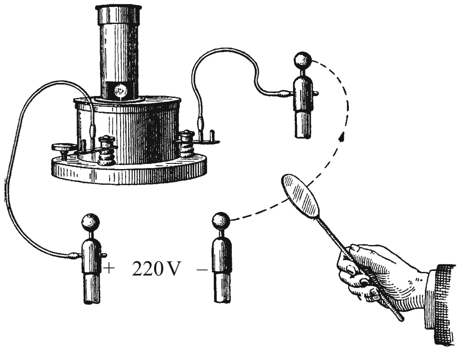
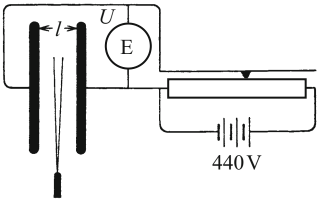


Figure 2.38 Measurement of the charge on a "carrier of electricity" (the galvanometer is the same as in Fig. 2.36)

2.12 The Electric Field

We follow the measurement of electric charges by measurements of the associated *electric field*. The principal characteristic of an electric field is its *preferred direction*, as indicated by the field lines. To describe an electric field quantitatively, we must therefore use a *vector*. We call it *the electric field* $\boldsymbol{E}$. The *direction* of this vector is the same as that of the field lines, conventionally from + to −. The *magnitude* of the vector can be determined by suitable experiments. One example can be carried out with the aid of two devices (Fig. 2.39):

Figure 2.39 The definition of the electric field strength



1. Flat parallel-plate condensers of differing plate areas A and plate separations l, and

2. Some sort of indicator for the electric field (a "field electroscope").

The *indicator* need only be able to identify two spatially or temporally separated electric fields as *equal*. It need not *measure* them, but rather only verify the *equality* of the two fields.

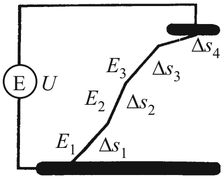
As an indicator, we choose two small[5], fine gold-covered quartz fibers, such as we have already seen in Fig. 2.4. We place them with their plane parallel to the field lines and observe the spread of their tips on a scale using an optical projection method[6].

During the experiments, we can vary the voltage between the condenser plates at will. To accomplish this, we use the familiar voltage divider circuit (Fig. 1.27). We will investigate a series of condensers with varying plate areas A and plate separations l. By regulating the voltage, we adjust the spread angle of the electroscope fibers to be the same each time; this means that the same force is acting on them and thus the same electric field is present each time. We find in this way a simple experimental result: The electric fields are the same as long as the ratio U/l, i.e. (voltage/plate spacing), is the same. The surface area of the plates plays no role. *The homogeneous electric field of a flat parallel-plate condenser is determined uniquely by the quotient* U/l. For this reason, one makes use of the ratio U/l to arrive at a first definition of the electric field strength (the magnitude of the

[5] Otherwise, they would distort the field too strongly; compare Fig. 2.25b.

[6] For thought experiments, a different indicator would be preferable, namely a tiny charged carrier of electricity attached to the arm of a force meter.

Figure 2.40 The path integral of the electric field $\boldsymbol{E}$



vector $\boldsymbol{E}$) in a parallel-plate condenser:

$$\text{Field strength } E = \frac{\text{Voltage } U \text{ between the condenser plates}}{\text{Spacing } l \text{ of the condenser plates}} . \qquad (2.2)$$

We use as unit 1 V/m.

The next step leads to an important generalization. By comparing with the homogeneous electric field of a parallel-plate condenser, one can measure the field $\boldsymbol{E}$ at arbitrary points within an arbitrary electric field: Its individual small regions are practically still homogeneous. They are replaced by the *field of the same strength and direction* of a parallel-plate condenser. We then determine for this substitute or "surrogate condenser" its field direction and the value of the ratio voltage/plate spacing (Eq. (2.2)).

The vector nature of the electric field $\boldsymbol{E}$ leads to a relation that is often useful: In Fig. 2.40, we have connected the arbitrarily-shaped conducting electrodes of a general condenser by a broken line. Within the individual path elements Δs along this line, the field must be practically homogeneous. Denote the components of the field in the direction of the path elements Δs as $E_1, E_2, \ldots, E_\mathrm{m}$. Then the sum $E_1\Delta s_1 + E_2\Delta s_2 + \ldots + E_\mathrm{m}\Delta s_\mathrm{m}$ is equal to $U_1 + U_2 + \ldots + U_\mathrm{m}$. We already know this latter sum, however: It is simply the voltage between the condenser plates. Therefore, we must have[C2.8]

$$\int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} = U \,, \qquad (2.3)$$

or, in words: The *path integral of the electric field along an arbitrary curve is equal to the voltage U between the ends of that curve.* We will make frequent use of this relation in the following pages.[C2.9]

The path integral changes its sign if it is computed in the opposite direction along the curve. It is positive when the curve is followed for the most part in the direction of the field, i.e. from + to − (compare Sect. 3.8).

In metrology, the measurement of electric field strengths plays a very minor role. *In the vast majority of cases, one* computes *the field strength E*. Examples can be found in Sect. 2.15. For the generally most important electric field, the homogeneous field of a flat parallel-plate condenser, this computation is dispensed with by referring to the defining equation (2.2).

C2.8. POHL for simplicity ignores the question of the sign here. Since he is dealing only with the *voltage U* between two points, it suffices here to use the *magnitude* of U. With the definition of the voltage as a *potential difference* $\Delta\varphi$ (Sect. 3.8), Eq. (2.3) becomes, taking the sign into account,

$$\Delta\varphi = -\int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} \,.$$

C2.9. The sentence in italics here is by no means trivial. It is also not true in general, as will be shown in Chap. 5. However, it holds for all the electric fields which are treated in the present chapter. They are characterized by the fact that their path integrals are independent of the curve chosen, and are therefore zero along any *closed* curve. Mathematically, such electric fields can be represented as the gradient of a scalar field (potential field). They are termed "conservative fields" or "potential fields".

2.13 The Proportionality of Charge Density and Electric Field Strength

In all of the electric fields that we have considered thus far, the field lines had ends, and electric charges were found to be sitting at those ends. Therefore, a quantitative relation between charge Q and electric field strength E is to be expected. We search for it experimentally by looking at the geometrically simplest field, our old friend the homogeneous field of a parallel-plate condenser. In Fig. 2.41, we see a condenser; let the area of each of its plates be A, the voltage between them be U, and their spacing l. Then the magnitude of the electric field between the plates is $E = U/l$, as seen above. In the figure, a *ballistic* galvanometer can be connected to the condenser. It has been ballistically calibrated and measures the current impulse $\int I\,\mathrm{d}t$ when the condenser is discharged (when contacts 1 and 2 are closed). We thus measure the magnitude of the two equal positive and negative charges Q which were on the condenser plates (e.g. in ampere second).

We repeat these measurements several times with various values of the plate area A and the field strength $E = U/l$. The result of all these measurements is

$$\frac{Q}{A} = \varepsilon_0 E\,, \tag{2.4}$$

or, in words: The surface density Q/A of the charge on the condenser plates is proportional to the electric field strength E (ε_0 is a constant of proportionality).

We find the same simple relation for the surface density Q'/A' of the *influence charges*. In Fig. 2.42, we see again the influence experi-

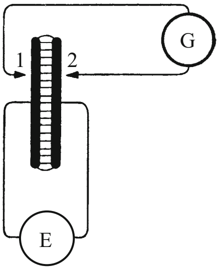


Figure 2.41 The proportionality of the field strength and the surface charge density

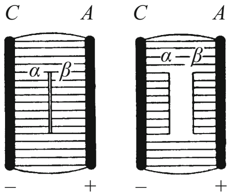


Figure 2.42 The measurement of the displacement density D as the surface density Q'/A' of the influence charge Q' ($U \approx 8000\,\mathrm{V}$)

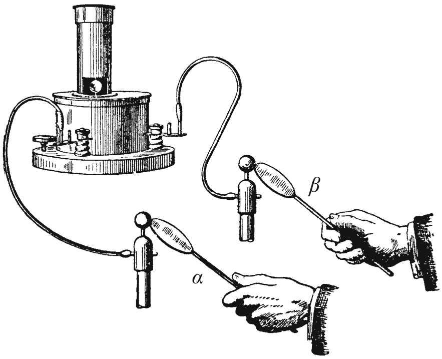


Figure 2.43 The influence charge Q' on the disks α and β is being measured as the "impulse deflection" of a ballistic galvanometer. Its calibration was carried out as in Fig. 1.33.

ment (Sect. 2.7) in a homogeneous field, now using rather thin metal disks α and β, with areas A', which cause hardly any distortion of the field. At the left in the figure, α and β are still in contact; at the right, they have been separated, still in the field. In Fig. 2.43, they have been taken out of the field and their charges Q' are being measured. The surface density of these influence charges has its own name; it is called the *displacement density* D, with $D = Q'/A'$. As its unit, we use for example $1\,\mathrm{A\,s/m^2}$. The displacement density $\boldsymbol{D}$ is also a *vector*. This can be seen by the fact that the influence charges depend on the inclination of the disks relative to the direction of $\boldsymbol{E}$. The largest induced (influence) charge is found when the disks are perpendicular to the field vector $\boldsymbol{E}$. This leads us to conclude that $\boldsymbol{D}$ lies parallel to $\boldsymbol{E}$ (Eq. (2.5)).

> The word "displacement" is unfortunate. It is intended to remind us of the displacement of the charges when the field is interrupted due to influence .[C2.10]

Inserting the displacement density $\boldsymbol{D}$, Eq. (2.4) takes on the form:[C2.11]

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} \,. \tag{2.5}$$

This is the essential content of the law discovered by CHARLES A. COULOMB in 1785. *It relates a charge density* (quoted e.g. in $\mathrm{A\,s/m^2}$), *measured via a current impulse, through a constant of proportionality* ε_0 *to an electric field E, measured in terms of a voltage* (quoted e.g. in V/m).

For the factor ε_0, in a vacuum (and for all practical purposes also in air), one finds the value:[C2.12]

$$\varepsilon_0 = 8.854 \cdot 10^{-12} \frac{\mathrm{A\,s}}{\mathrm{V\,m}} \,.$$

It is called the *electric field constant* or the *permittivity of vacuum*; its official (SI) name is simply the *electric constant*.

> For precise measurements of the electric field constant, instead of the simple condenser sketched in Fig. 2.41, one makes use of a condenser with

C2.10. POHL suggests in the 21st edition that one could simply use the expression "electric field quantity $\boldsymbol{D}$" instead of "displacement". $\boldsymbol{D}$ is also called the "electric flux density".

C2.11. Equation (2.5) or (2.4) is a special case of the general relation (2.9) which is discussed in the following section:

$$\frac{Q}{\varepsilon_0} = \oint \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{A} \,.$$

It is the first of the MAXWELL equations in integral form, and is also known as GAUSS's law (or GAUSS's formulation of COULOMB's law). The connection to COULOMB's law (as given in Eq. (3.8)) can be seen if we compare the equation for the field strength of a charged sphere (Eq. 2.15) with that equation and use the expression for the force on a charge in an electric field,

$\boldsymbol{F} = Q\boldsymbol{E}$ (Eq. (3.5)).

(The differential form of the first MAXWELL equation can be found at the end of Sect. 2.14.)

For the history of the discovery of COULOMB's law, see e.g. J. L. Heilbron, p. 470, cited in the footnote in Sect. 2.17.

C2.12. When the quantities current and voltage are considered to be independent of each other, as we do here, then ε_0 is a natural constant which must be determined experimentally. But we should mention that the value of ε_0 has been fixed by law today (see Comment C23.1).

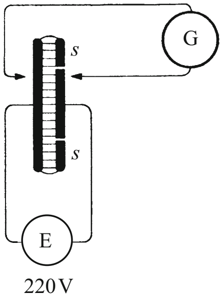


Figure 2.44 The same experiment as in Fig. 2.41, but using a condenser with a corona ring

a "corona ring" S (Fig. 2.44). The surface charge density is measured only on the inner part of the condenser, thus avoiding disturbances due to the inhomogeneous stray field around the edges of the plates.

The *experimental fact* summarized in Eq. (2.5) can be interpreted in three ways:

1. One considers the readily-measurable quantity $\boldsymbol{D}$ as a useful aid to the measurement of electric fields $\boldsymbol{E}$, employing $\boldsymbol{E} = \boldsymbol{D}/\varepsilon_0$;

2. One can treat the displacement density $\boldsymbol{D}$ simply as an abbreviation for the product $\varepsilon_0\boldsymbol{E}$, which occurs frequently in electromagnetism;

3. or, one can treat $\boldsymbol{D}$ as an independent, second quantity which is equivalent to $\boldsymbol{E}$ for the quantitative description of electric fields.[C2.13]

In this book, we will place all three of these possibilities on an equal footing.

C2.13. The introduction of *two* vector fields, $\boldsymbol{E}$ and $\boldsymbol{D}$, which differ only through a constant of proportionality ε_0, might seem unnecessary here. For the complete description of an electric field in vacuum, $\boldsymbol{E}$ is indeed sufficient. $\boldsymbol{D}$ is therefore not even mentioned in some textbooks. However, in the presence of dielectric materials, the simple relation $\boldsymbol{D} = \varepsilon_0\boldsymbol{E}$ no longer holds. Then it is often found to be helpful to employ both of the field quantities (see Chap. 13).

2.14 The Electric Field of the Earth. Space Charge and Field Gradients. MAXWELL's First Law

The earth is always surrounded by an electric field $\boldsymbol{E}$ (G. LE MONNIER, physician, 1752). This field points downwards and is perpendicular to the surface of the earth in flat regions. The field can be readily detected and measured by making use of Eq. (2.5). We employ a flat parallel-plate condenser which can be rotated around its horizontal axis (Fig. 2.45). It is set up in the open. Each plate (made for example of aluminum sheet metal) has an area A of about $1\,\mathrm{m}^2$. The plates are analogous to the small disks in the influence experiment. They are each connected to one terminal of a ballistic galvanometer which is calibrated in ampere second. We alternately orient the plane of the plates in the vertical and the horizontal, that is alternately parallel and perpendicular to the field. With each change of orientation, the galvanometer indicates a current impulse $Q =$

Video 2.3:
"Measuring the electric field of the earth"
http://tiny.cc/69fgoy
In the experiment, the condenser is rotated by 180°, so that the current impulse is doubled.

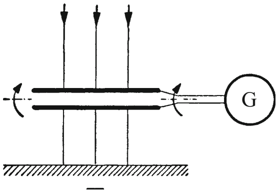


Figure 2.45 A measurement of the displacement density of the electric field of the earth using a rotatable parallel-plate condenser connected to a ballistic galvanometer **(Video 2.3)**

$\int I\,\mathrm{d}t$ of around 10^{-9} A s. The ratio Q/A is the displacement density D of the electric field of the earth. Averaged over time, one finds

$$D = 1.15 \cdot 10^{-9}\,\frac{\mathrm{A\,s}}{\mathrm{m}^2}$$

or

$$E = \frac{D}{\varepsilon_0} = 130\,\frac{\mathrm{V}}{\mathrm{m}}\,.$$

The earth has a surface area A_e of $5.1 \cdot 10^{14}\,\mathrm{m}^2$. Its total negative charge is thus $A_\mathrm{e} \cdot D \approx 6 \cdot 10^5$ A s. Where are the corresponding positive charges? One could think of the fixed stars; in that case, we would be dealing with the usual radially-symmetric electric field of a charged sphere at a great distance from other objects or charges (Fig. 2.48). The electric field strength would be practically the same at several kilometers altitude above the earth's surface as on the surface itself (the earth's radius is 6370 km!). But this is not at all the case. Already at an altitude of 1 km, the field strength has dropped to around 40 V/m. At an altitude of 10 km, it is only a few V/m.

These observations lead us to a new kind of electric field and thereby to a fundamental relation between charge and electric field. The fields which we have treated thus far were delimited on both sides by a solid body which carried electric charges. In the case of the earth's field, we have a solid body on only one side, namely the earth itself as carrier of the negative charges. The corresponding positive charges are localized on innumerable, invisible carriers in the atmosphere. These carriers all together form a cloud of positive *space charge* (Fig. 2.46). The volume density ϱ of these charges (A s/m^3) is responsible for the strong "slope" (the gradient) of the field. We find

$$\varrho = \frac{\partial D}{\partial x} = \varepsilon_0 \frac{\partial E}{\partial x}\,. \tag{2.6}$$

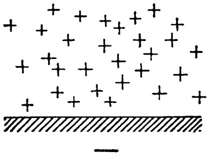
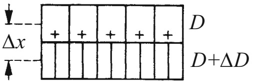


Figure 2.46 A cloud of positive space charge above the negatively-charged surface of the earth (approximated here as a planar surface)

Figure 2.47 The relation between a field gradient and a space charge



Derivation: In Fig. 2.47, two homogeneous field regions are sketched, with the cross-sectional area A and displacement densities D and $(D + \Delta D)$, one above the other. D is thus assumed to increase by the amount ΔD on passing down the vertical path element Δx. Then from Eq. (2.4), it follows that

$$\varepsilon_0 \Delta E = \Delta D = \frac{\Delta Q}{A} \tag{2.7}$$

or

$$\varepsilon_0 \frac{\Delta E}{\Delta x} = \frac{\Delta D}{\Delta x} = \frac{\Delta Q}{A \Delta x} = \varrho\,, \tag{2.6}$$

since ΔQ is the charge contained within the volume $A\,\Delta x$. It is marked by the + signs in Fig. 2.47.

Equation (2.6) is a special case, limited to a gradient in *one* direction (along the x axis); usually, this relation is written in the general mathematical form (obtained in the limit $\Delta x \to 0$ and in three dimensions):

$$\operatorname{div} \boldsymbol{E} = \frac{\varrho}{\varepsilon_0} \tag{2.8}$$

and is called the “fundamental equation of electrostatics”. It is one of the four MAXWELL *equations* (see Sect. 6.5), and it describes the general relation between the vector field $\boldsymbol{E}$ and the charge density ϱ (**Exercise 2.12**). In integral form, it is given by

$$\oint \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{A} = \frac{1}{\varepsilon_0} Q\,, \tag{2.9}$$

where the (2 dimensional or surface) integral is to be carried out over the closed surface area $\boldsymbol{A}$ which encloses the charge Q. The surface integral $\int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{A}$ is also called the *electric flux*.

2.15 The Capacitance of Condensers and Its Calculation

Combining the two equations

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} \tag{2.5}$$

and

$$\int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} = U\,, \tag{2.3}$$

we can compute the distribution of the electric field strength of arbitrary fields. We thus arrive at the physically as well as technologically important concept of the *capacitance*. The capacitance is defined for every condenser as the quotient

$$C = \frac{\text{Charge } Q \text{ on the electrodes}}{\text{Voltage } U \text{ between the electrodes}} . \tag{2.10}$$

Its unit is 1 ampere second/volt (A s/V), and is given the name 'farad' (F) (after Michael FARADAY).

Q is the electric charge on one of the electrodes (which define the boundaries of the electric field), and has the same magnitude as the charge (of opposite sign) on the other electrode. One is positive and the other negative. Often, we speak conveniently, but less rigorously, of the "charge on the condenser", and correspondingly of "charging" and "discharging" the condenser (this will be treated quantitatively in the following section).

We give examples of the capacitance of several condenser designs with geometrically simple electric fields:

1. A *flat parallel-plate condenser*. In its homogeneous field, the displacement density D is equal to the surface charge density Q/A of the charge on the two condenser plates (electrodes). Equation (2.2) then gives the field strength $E = U/l$. Inserting both into Eq. (2.5) gives

$$C = \varepsilon_0 \frac{A}{l} . \tag{2.11}$$

A *numerical example*: 2 circular plates of 20 cm diameter and areas of $3.14 \cdot 10^{-2}\,\text{m}^2$, at a spacing of 4 mm:

$$C = \frac{8.86 \cdot 10^{-12}\,\text{A s} \cdot 3.14 \cdot 10^{-2}\,\text{m}^2}{\text{V m} \cdot 4 \cdot 10^{-3}\,\text{m}} = 7 \cdot 10^{-11}\,\frac{\text{A s}}{\text{V}}$$
$$(\text{or } 7 \cdot 10^{-11}\,\text{farad} = 70\,\text{pF}) .$$

Here, as in Sect. 2.8 for several resistors, it is worth the trouble to derive the capacitance of several condensers connected in series or parallel. We readily obtain for two condensers in parallel$^{\text{C2.14}}$

$$C = C_1 + C_2 . \tag{2.12}$$

In a series circuit, we find

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} . \tag{2.13}$$

2. A *spherical condenser electrode with a radius r* (and a radially-symmetric electric field; see Fig. 2.48). A charge Q is located on the surface of the sphere. At a distance R from the center of the spherical electrode, it gives rise to a displacement density$^{\text{C2.15}}$

$$D_{\text{R}} = \frac{Q}{4\pi R^2} , \tag{2.14}$$

C2.14. When two condensers are connected in parallel, their two negative electrodes are equivalent to a larger electrode whose total area is the sum of the two individual electrode areas, and similarly for the positive electrodes. They thus act like a larger condenser, and the overall capacitance of the circuit is just the sum of the individual capacitances of the two condensers.
In a series circuit, the voltages on the individual condensers simply add, and since the capacitance is proportional to the reciprocal of the voltage, the reciprocal of the overall capacitance is just the sum of the reciprocals of the individual capacitances; this is analogous to resistors connected in parallel.

C2.15. One can try to derive Eq. (2.14) by imagining that pairs of concentric spherical surfaces are brought into the field, and then computing the density of surface influence charges (displacement densities) which would be present on them. For those who prefer an experimental derivation, we refer to Comment C2.16.

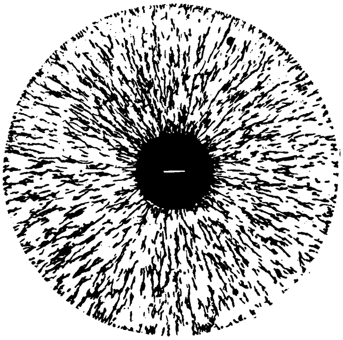


Figure 2.48 The radially-symmetric electric field lines between a negatively-charged sphere and very distant positive charges

and, from Eq. (2.5), the field strength$^{C2.16}$

$$E_R = \frac{Q}{4\pi\varepsilon_0 R^2}\,. \tag{2.15}$$

The voltage U between the charged sphere and the very distant boundary of the field (e.g. the walls of the room) is obtained from Eq. (2.3) by integration. Then

$$U = \int\limits_{R=r}^{R=\infty} E_R\,\mathrm{d}R = \int\limits_{R=r}^{R=\infty} \frac{Q\,\mathrm{d}R}{4\pi\varepsilon_0 R^2} = \frac{Q}{4\pi\varepsilon_0 r}\,. \tag{2.16}$$

Equations (2.10) and (2.16) together yield the capacitance of a spherical electrode (spherical condenser):

$$C = 4\pi\varepsilon_0 r \tag{2.17}$$

$$\left(4\pi\varepsilon_0 = 1.11\cdot 10^{-10}\,\mathrm{A\,s/V\,m}\right).$$

The capacitance of a sphere is proportional to its radius.

In Fig. 2.49, we measure the capacitance C of a globe which has been hung on an insulating cord to verify Eq. (2.17). A voltage of 220 V suffices for this measurement.

> The earth's radius is $r = 6.37\cdot 10^6$ m. It thus forms with the system of fixed stars (as counter electrode) a condenser whose capacitance according to Eq. (2.17) is 708 microfarad (µF).

In an analogous manner, we can compute the electric fields of more complex electrode shapes, as long as they are sufficiently symmetric, as well as the spatial distribution of the field strength and their capacitances[7]

C2.16. The experimental derivation of the important equation (2.15) starts with the experiment illustrated in Fig. 2.49 **(Video 2.4)**, which can be carried out with spheres of different radii. Measurement of their capacitances $C(r)$ as a function of the radius r of the sphere yields Eq. (2.17), and

$$U = \frac{Q}{4\pi\varepsilon_0 r} \quad (2.16)\,.$$

Substituting Eq. (2.15), it follows that

$$U = \int\limits_{R=r}^{R=\infty} E_R\,\mathrm{d}R$$

(concerning the sign, see Comment C2.8). The importance of this equation is that it is experimentally verified in principle for all radii, including extremely small values of r ("point charges"). Thus, by superposition of many such point charges, the electric fields and potentials (Chap. 3) of arbitrary known charge distributions can be calculated.

[7] Examples:

$$\text{2 concentric spheres: } C = 4\pi\varepsilon_0\frac{r_1 r_2}{r_2 - r_1}\,, \tag{2.18}$$

$$\text{2 coaxial cylinders of length } a\text{: } C = 2\pi\varepsilon_0\frac{a}{\ln(r_2/r_1)}\,. \tag{2.19}$$

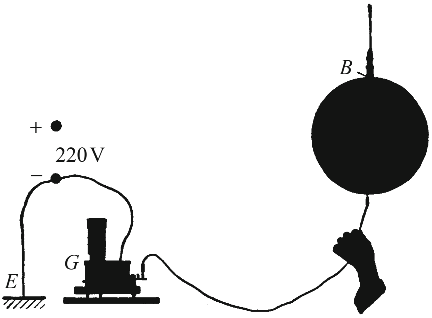


Figure 2.49 Measurement of the capacitance of a condenser formed by a sphere and the floor of the lecture room. To charge it, the cardboard sphere is connected briefly to the + pole of the current source ($U = 220\,\text{V}$). The negative pole of the current source was previously connected to the earth E (it was "grounded"). (The calibration of the ballistic galvanometer G in ampere second was carried out as in Fig. 1.33; B: insulator) **(Video 2.4)**

Video 2.4: "The capacitance of a sphere"
http://tiny.cc/jaggoy
The globe (of radius $r = 0.27\,\text{m}$) is charged up to a voltage of 10^3 V. The current impulse on discharging it produces a galvanometer deflection of 3.7 scale divisions in the ballistic galvanometer G, corresponding to $4 \cdot 10^{-8}$ A s. Doubling the charging voltage doubles the measured charge. The capacitance is found to be $C = 4 \cdot 10^{-11}$ farad. The capacitance calculated from Eq. (2.17) is $3 \cdot 10^{-11}$ farad.

For an overview of complex fields, we give a useful tip: The combination of Eqns. (2.15) and (2.16) yields the field strength directly on the surface of the sphere (where $R = r$!):

$$E_\text{r} = \frac{U}{r}\,. \tag{2.20}$$

Any sharp corner or point can to first order by considered as a spherical surface with a small radius of curvature r. From Eq. (2.20), the field strength E on the surface of a sphere and the radius of curvature r of the sphere are inversely proportional to each other. Therefore, in the neighborhood of corners and points of condenser electrodes, even quite low voltages give rise to high field strengths. The air becomes conducting at high field strengths ("field ionization" of air molecules) and is no longer an insulator. A violet-colored glow ("corona discharge") indicates fundamental changes in the air molecules. In addition, an *electric wind* is produced: It blows away from the point and is a first example of the *matter transport* which is associated with an electric current.

The air which streams away from the point is replaced by air which comes in from the sides; it is in turn ionized and accelerated away from the point. A counter-force acts on the point. It can for example cause the "propeller" sketched in Fig. 2.50 to rotate. The voltage between the propeller and the walls of the room need be only a few thousand volt .[C2.17]
Apart from details, the same process occurs as with a *propeller aircraft*: The propeller or fan accelerates the air which enters from the sides and blows it backwards as a jet. The counter-force which is oppositely directed to the jet of air accelerates the aircraft on takeoff and later allows it to maintain a constant velocity in the face of the inevitable frictional resistance to its motion (Vol. 1, Sects. 5.11 and 10.11).

C2.17. A planned application of this matter transport is a rocket drive for space flight using ion beams.

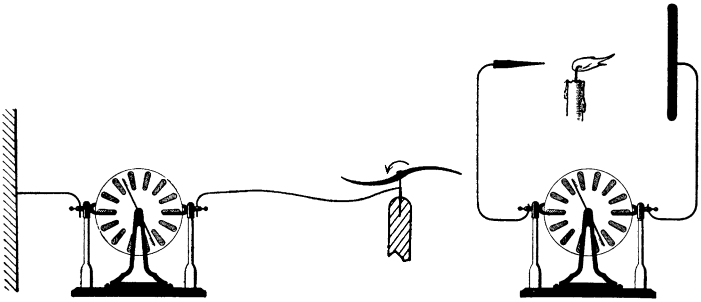


Figure 2.50 *At left*: Pinwheel (ANDREAS GORDON, 1712–1751) (**Video 2.5**).

Video 2.5:
"The Electric Wind"
http://tiny.cc/maggoy
Right: Ionic wind (particularly effective as a shadow image). – An instructive variation: Hang a light-weight condenser made of a pointed and a ring-shaped electrode mounted rigidly, with two thin wires which serve as electrical leads and suspension. This "pendulum" swings whenever the jet of the electric wind is blowing through the ring.

2.16 Charging and Discharging a Condenser

The time dependence of discharging a condenser was already shown in Fig. 2.33. In order to investigate it quantitatively, we use the circuit shown at the upper right in Fig. 2.51. As can be seen there, the total voltage is the sum of the condenser voltage and the resistor voltage:

$$U = U_\mathrm{C} + U_\mathrm{R} = \frac{Q}{C} + RI\,, \qquad (2.21)$$

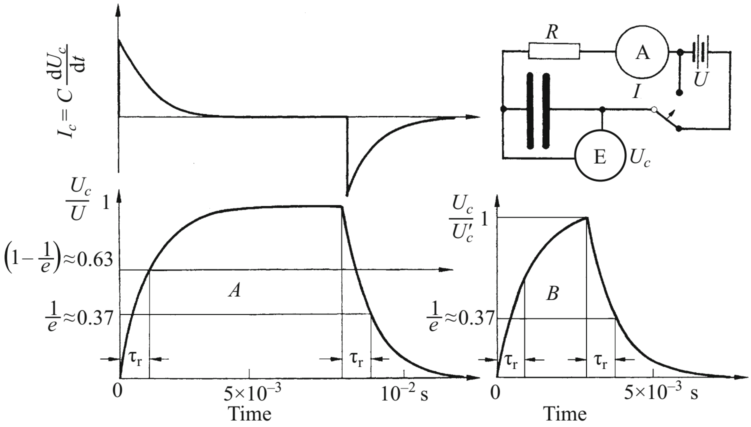


Figure 2.51 Production and decay of the electric field in a condenser as a function of time, i.e. charging and discharging the condenser. An oscilloscope is used here to measure the voltage (shown as a static voltmeter in the circuit at upper right). From its time dependence (curve *A*), we obtain by differentiation the time dependence of the current shown above (Eqns. (2.23) and (2.24)). *B*: Discharging begins already at $U'_\mathrm{c} < U$. ($C = 10^{-6}$ F, $R = 10^{3}\ \Omega$, τ_r = relaxation time = $RC = 10^{-3}$ s)

and thus,

$$U = \frac{Q}{C} + R\frac{\mathrm{d}Q}{\mathrm{d}t}\,. \tag{2.22}$$

This differential equation has the solutions

$$Q = Q_0\,\mathrm{e}^{-\frac{t}{RC}} \quad \text{or} \quad I = -\frac{Q_0}{RC}\,\mathrm{e}^{-\frac{t}{RC}} \tag{2.23}$$

on discharging ($U = 0$),[C2.18] and

$$Q = Q_0\left(1 - \mathrm{e}^{-\frac{t}{RC}}\right) \quad \text{or} \quad I = \frac{Q_0}{RC}\,\mathrm{e}^{-\frac{t}{RC}} \tag{2.24}$$

for charging ($U = U_0$), as one can verify by substituting into Eq. (2.22) (simple differential equations can best be solved by following this recipe: Guess the solution and test it by substituting into the original equation). The time $\tau_\mathrm{r} = RC$ is called the *relaxation time* or *time constant* of the circuit. It can be used to measure resistances whose value is greater than the order of magnitude of $10^7\,\Omega$.

C2.18. This result clearly does not hold for $R = 0$, since then the condenser would be discharged within an infinitely short time, without the energy it contains being converted into JOULE heat. The current in this case would be determined by other properties of the circuit which are not taken into account in Eq. (2.22), and which would lead to oscillations (see Sect. 11.2).

2.17 Various Types of Condensers. Dielectrics and Their Polarization

Thus far, we have used condensers of practically only two types. They consisted either of a pair of parallel plates (Fig. 2.1), or of several pairs of plates (Fig. 2.52). A variation on this multi-plate condenser is the rotary variable condenser (Fig. 2.53). By rotating

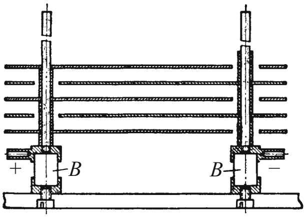


Figure 2.52 The design of multi-plate condensers. Usually, they have three instead of the one pair of mounting shafts shown here. (*B* are insulators)

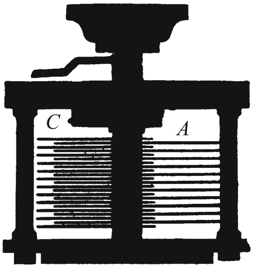


Figure 2.53 Shadow image of a rotary variable condenser

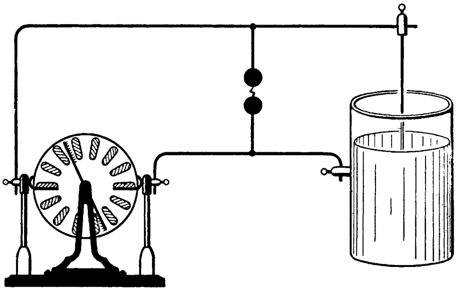


Figure 2.54 Charging a Leyden jar

one set of plates so that different fractions of the plate areas overlap, one can vary the capacitance of the condenser.

Technical condensers often have liquid or solid materials ("dielectrics") between their plates rather than air. We mention here three types which are frequently used:

1. The well-known and venerable *Leyden jar*[8]. Figure 2.54 shows at the right a primitive version: A glass cylinder has a sheet of tinfoil glued inside and another outside. Its capacitance is usually in the range of 10^{-9} to 10^{-8} F (1 to 10 nF).

A small influence machine produces currents of around 10^{-5} A (Sect. 1.9). With this current, it can charge a Leyden jar with a capacitance of 10^{-8} F in 30 s to a voltage of about $3 \cdot 10^4$ V (Fig. 2.54). A spark gap with two balls at a distance of 1 cm, wired in parallel to the Leyden jar, can serve as a rough voltmeter. At around 30 000 V, a spark jumps the gap, accompanied by a loud 'snap'. The time during which a spark of this kind jumps is ca. 10^{-6} s. This can be registered with a rapidly-rotating photographic plate (or with a photo detector and an oscilloscope). The current in the spark must therefore be $30/10^{-6}$ or $3 \cdot 10^7$ times larger than the current from the influence machine; it must be around 300 A. This large current causes a strong heating of the air in the spark gap, with the resulting noise (thunderclap principle).

2. The *paper condenser*. We lay out two strips of metal foil C and A, separated and covered by two strips of insulating paper P,P, then roll up the whole packet and press it together (Fig. 2.55). Newer variants use plastic foils with vapor-deposited metal electrodes on each side. They can be fabricated for voltages of up to several thousand volt.

3. *Electrolytic condensers*. In these, the insulating separation layer (dielectric layer) is produced electrolytically and has a thickness of the order of 0.1 μm. Condensers of capacitances up to 10^{-3} F, or even 1 F are commercially available today, and can be used at up to several hundred volt.

The treatment in this and the following chapter is limited to electric fields in *vacuum*, which is practically the same as in air. *Matter* in an electric field will be discussed in Chap. 13. Nevertheless, in the three

[8] See: J.L. Heilbron, "*Electricity in the 17th and 18th Centuries: A study of early modern physics*", University of California Press, Berkeley, CA (1979), p. 309.

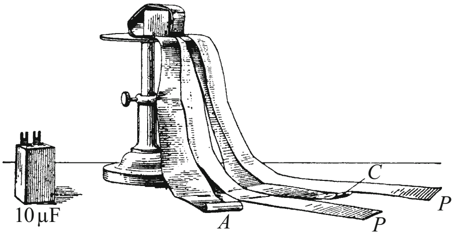


Figure 2.55 *At the left*: A finished paper condenser with a capacitance of 10 µF; *at the right*: The condenser is partially unrolled. The two metal foils each have an area of about 4 m^2. Their spacing, equal to the thickness of the paper strips P, is about 0.02 mm (**Exercise 2.13**).

types of condenser described above, we have intentionally anticipated that topic. We therefore introduce three new concepts at this point, the *dielectric*, its *polarization*, and its *dielectric constant*.

A good insulator causes an electric field to decay only very slowly. It can be "penetrated" by an electric field over a long period of time: Thus its name, "dielectric".

The ratio

$$\varepsilon = \frac{\begin{array}{c}\text{Capacitance } C_\text{m} \text{ of a condenser}\\ \text{completely filled with a dielectric}\end{array}}{\text{Capacitance } C_0 \text{ of the empty condenser}} \tag{2.25}$$

is called the *dielectric constant* of the dielectric. Numerical values will be given in Table 13.1.

With a given charge on the electrodes of the condenser, an increase in its capacitance is accompanied by a decrease in its voltage. Filling the gap between its electrodes with a dielectric thus produces a similar effect to partially filling the gap by a conducting object (Fig. 2.3). The conductor cancels the field in its interior. It thereby shortens the field lines by the amount of its thickness. At the same time, charges are induced on its surfaces: this is the phenomenon of influence.

In an insulator or dielectric, the charges cannot move to the surface as they do in a metal. Nevertheless, an insulator in an electric field causes a shortening of the field lines; in the simplest case, one can simply assume that influence is operating on the scale of the individual molecules. This is illustrated in Fig. 2.56 by a rough two-dimensional model. The molecules are represented arbitrarily by small conducting spheres. Influence within individual molecules is called an *electric polarization of the molecules*. It in turn produces a "polarization of the whole dielectric". Then charges appear on its surfaces, just as with influence in conductors; in Fig. 2.56, positive on the left and negative on the right. But the polarization of an insulator cannot be used to separate charges like the influence in a conductor.

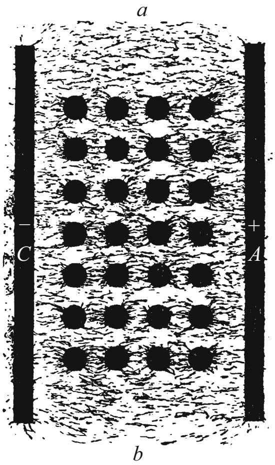


Figure 2.56 A model experiment to explain the polarization of a dielectric in terms of polarization of its individual molecules (Exercise 2.14)

Imagine that the *polarized* insulator in Fig. 2.56 is split into two parts along the cross-section $a\,b$ perpendicular to the field lines and the two halves are removed from the field: Then each half contains the same number of $+$ and $-$ charges, and is thus wholly uncharged. The polarization charges are also called "bound" charges, to distinguish them from the "free" charges on the condenser plates or the surfaces of conductors in the field.

The model experiment in Fig. 2.56 contains extensive but not essential simplifications. In reality, the molecules are not spherical, and the charges do not move to the ends of the molecules. More details are to be found in Sect. 13.9. In any case, a rather unspectacular experiment, sliding an insulator into the gap between the plates of a condenser (Fig. 2.3), has led us to an important result: *In the interior of matter, charges are present. They can be displaced by an external electric field. This results in an "electric deformation" or polarization of the molecules themselves.*

What happens when an object is electrically charged according to this picture? From Sect. 2.7, this can only mean that an *excess* of charges of one sign is present on the object. But how many charges of *both signs* are present, whose *difference* is observed in charged bodies? We offer an example:

An amount of water of mass $M = 1\,\text{kg}$ has a volume of $V = 10^{-3}\,\text{m}^3$ and, if it were spherical, a radius $r = 6.2\,\text{cm}$. From the molar mass of the water molecules, $M/n = 18\,\text{g/mol}$, we find the amount of substance in the sphere to be $n = 55.56\,\text{mol}$, and thus the number N of water molecules to be $N = n\,N_\text{A} = 3.34{\cdot}10^{25}$ molecules, each with 10 electrons. Since the charge of a single electron is $1.602 \cdot 10^{-19}\,\text{A s}$ (Sect. 3.6), the sphere contains charges Q of both signs, each totalling $5.4 \cdot 10^7\,\text{A s}$. Between this sphere and the walls of the laboratory, we could – with some difficulty – produce voltages of $U > 10^6\,\text{V}$. Then, from Eq. (2.17), a charge $q = 6.9 \cdot 10^{-6}\,\text{A s}$ would sit on the surface of the sphere, either positive or negative. The ratio q/Q is then $1.3 \cdot 10^{-13}$. This means that, although we would say in the laboratory that a large electric charge was present on the object, in reality we would have removed or added only an unimaginably small fraction of its positive or negative charges Q and thereby produced a difference of $q = Q_+ - Q_-$ (that is, we have disturbed the electric equilibrium in the object, but only by a very small amount). Only when the object has an exceedingly small mass, i.e. for individual molecules or atoms, can the charge difference q be of the order of the total charge Q.

Exercises

2.1 For a simple application of Eq. (2.8), we assume that the cloud of positive space charges in Fig. 2.46 has a constant volume charge density ϱ up to a height of $x = h$ and then $\varrho = 0$ for $x > h$. Let the negative surface charge density σ be that of the earth's surface, and h the height, both known. Then, find a) the volume charge density ϱ; b) the electric field E as a function of the height x above the earth's surface; c) the voltage U_x between the earth and a point at x above it; and d) the voltage U_h between $x = h$ and $x = 0$ (Sect. 2.14).

2.2 In the experiment shown in Fig. 2.35, the electrolysis of water is carried out by discharging a condenser. How large must the capacitance C of the condenser be, if it were charged up to a voltage of 220 V and produced 10 mm^3 of hydrogen gas on discharging through the electrolysis cell? The number density N_V of the H_2 molecules at 300 K and atmospheric pressure is $N_\mathrm{V} = 2.45 \cdot 10^{25}\,\mathrm{m}^{-3}$ (Sect. 2.15, see also Sect. 1.4).

2.3 Two condensers with the capacitances C_1 and C_2 are connected in series. Find their overall capacitance C for the case that $C_1 = 1$ nF (a small Leyden jar) and $C_2 = 2\,\mu$F (Sect. 2.15).

2.4 An uncharged parallel-plate condenser with a capacitance of $C = 0.1$ nF is connected to a static voltmeter which was previously charged and indicated a voltage U. When the condenser is connected, the voltage decreases by 10 %. Find the capacitance C_V of the voltmeter (Sect. 2.15).

2.5 The voltage U between a negatively-charged sphere of radius $r = 1$ cm and the distant walls of the room is $U = 10^5$ V (see Video 2.3). Determine the surface charge density at the surface of the sphere, expressed as the surface number density N_e of electrons (Sect. 2.15).

2.6 A coaxial cable of length l has an inner conductor of diameter $2r = 2\,\mathrm{mm}$ in a 5 mm thick plastic insulation with the dielectric constant $\varepsilon = 3$, which carries the outer conductor. Find the capacitance per length unit, C/l (Sects. 2.15, 2.17).

2.7 Two point charges $Q_1 = Q$ and $Q_2 = -3Q$ are at a distance of 1 m. At what point along their connecting line is a) the electric field strength $E = 0$, and b) the potential $\varphi = 0$ (besides at infinity)? (Sects. 2.15, 3.8).

2.8 A rotating-coil galvanometer (Fig. 1.19) has a resistance of $100\,\Omega$. When a charged condenser (capacitance $0.1\,\mu\mathrm{F}$, voltage 10 V) is discharged through this galvanometer, it shows a ballistic deflection of 10 scale divisions. Find the voltage impulse $\int U\mathrm{d}t$ which would produce a deflection of one scale division (Sect. 2.16).

2.9 A condenser of capacitance C is discharged through a resistor of resistance R. a) Determine the time $t_{1/2}$ in which the voltage decreases to one-half of its initial value; b) how large is the capacitance C of the condenser if $R = 10^{12}\,\Omega$ and the voltage decreases by 20 % after 10 seconds? (Sect. 2.16).

2.10 A neon lamp (Fig. 1.16) lights up at a voltage between 160 and 220 V; the current increases linearly with voltage over this range, following the equation $I = (U - 157\,\mathrm{V})/7.63\,\mathrm{k}\Omega$ (the "differential resistance" $\mathrm{d}U/\mathrm{d}I$ is thus $7.63\,\mathrm{k}\Omega$). The lamp is connected to a condenser charged to a voltage of 220 V; its capacitance is $C = 50\,\mu\mathrm{F}$. Find the time τ that it takes for the lamp to go out (Sect. 2.16).

2.11 A condenser (C), an OHMic resistor (R), a battery (U_0) and a switch are connected in series in a circular circuit. The switch is closed, so that the condenser is charged through the resistor up to the voltage U_0. Find the energy W_{batt} supplied by the battery and compare it with the energy W_{C} stored in the condenser. What role is played by the resistor R? (Sects. 2.16, 3.7).

2.12 A rotary condenser (Fig. 2.53) has semicircular plates with a radius of 5 cm and spacings of 1.1 mm. How many plates will be needed to obtain a maximum capacitance of 500 pF ? (Sect. 2.17).

2.13 Determine the dielectric constant ε of the paper strip P in Fig. 2.55, if for the dimensions given there, the capacitance obtained is $C = 10\,\mu\mathrm{F}$ (Sect. 2.17).

2.14 Find the effective dielectric constant ε in the model experiment illustrated in Fig. 2.56, in which metal balls of diameter d are arranged on a cubic lattice with a lattice constant $2d$ within the otherwise empty space between the condenser plates (Sect. 2.17).

2.15 A parallel-plate condenser (with rectangular plates of height a, width b, and spacing l) is dipped vertically into a liquid (density ϱ, dielectric constant ε) so that the lower edges of its plates just touch the liquid. When a voltage U is applied to the plates, the liquid rises between them to a height h, so that a volume $h \cdot l \cdot b$ is filled by the liquid. Find the height h (Sects. 2.17, 3.7).

2.16 The volume between the plates of a parallel-plate condenser ($V = A \cdot l$, $A =$ surface area and $l =$ spacing of the plates) is assumed to be completely filled with a liquid of dielectric constant ε. How large is the force with which the plates attract each other when a voltage U is applied? (Sects. 2.17, 3.7).

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_2) contains supplementary material, which is available to authorized users.

Forces and Energy in Electric Fields

3

3.1 Three Preliminary Remarks

1. In every physics laboratory for research or teaching, one can find an assortment of instruments for measuring time, length, mass, and temperature as well as for electric current, voltage, capacitance and various other electric quantities. However, *force meters* are to be found, if at all, only rarely and then mostly in teaching labs. If an investigation requires the measurement of a force, one generally compares that force to the force that we call *weight* (unit newton). In general, forces are not measured directly, but rather are computed from other quantities.

2. The relation between force $\boldsymbol{F}$, mass m, and acceleration $\boldsymbol{a}$ has to be derived experimentally. This problem is one of the most invidious in all of physics teaching. One method is described in detail in Vol. 1, Sect. 3.2. The corresponding experiments yield the result

$$\boldsymbol{a} = \frac{\boldsymbol{F}}{m} \quad \text{or} \quad \boldsymbol{F} = m\boldsymbol{a}$$

with a barely acceptable precision. The real justification for this fundamental equation is to be found only later in the successes of its numerous applications.

3. Exactly the same situation is found in electromagnetism for the fundamental equation derived in Sects. 3.2 and 3.3:

$$\boldsymbol{E} = \frac{\boldsymbol{F}}{Q} \quad \text{or} \quad \boldsymbol{F} = Q\boldsymbol{E}\,,$$

giving the relation between the mechanical quantity 'force' $\boldsymbol{F}$ and electrical quantities (charge Q, electric field $\boldsymbol{E}$). Again, for this equation, the final justification is found only later in terms of all of its general applications.

3.2 The Fundamental Experiment

We start, as always, from experimental results. Figure 3.1 shows a disk-shaped charge carrier α on the lever of a force meter, in this case a small beam balance. The carrier is placed at the center of

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_3

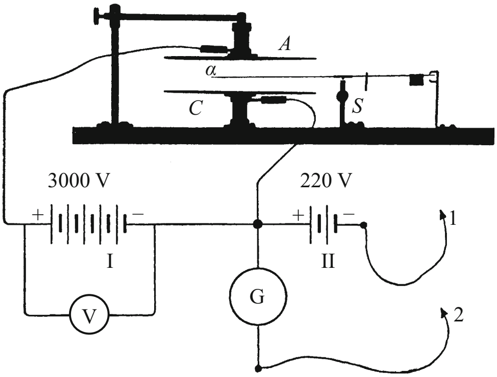
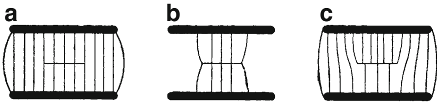


Figure 3.1 The fundamental experiment on the force between a charge and an electric field. The (insulating) quartz balance beam carries two riders made of Al sheet metal on its right side and can swing between two limiting stops. *S* is a small round weight which keeps the center of gravity of the balance beam below the knife-edge bearing. The condenser plates (electrodes) *A* and *C* are mounted on insulating posts.

Figure 3.2 The field-line patterns in the fundamental experiment of Fig. 3.1



a parallel-plate condenser, between the two electrodes *C* and *A*. Its shape and its orientation perpendicular to the field lines have been chosen for a reason: *The carrier, when itself uncharged, should have only a minimal effect on the form of the electric field between C and A* (Fig. 3.2a); *it should not distort the field by influence* (Fig. 2.25b). The field between *C* and *A* is produced with the aid of a current source *I*. It operates at the voltage *U*, so that the field strength of the homogeneous field in the condenser is $E = U/l$ (l = spacing of the electrodes). With this setup, we proceed as follows:

1. We transfer a negative charge to the carrier α. To do this, we connect it briefly to the negative pole (contact 1) and *both* condenser plates to the positive pole of the current source *I* ($U = 0$). After charging the carrier, we will observe the field pattern shown in part b of Fig. 3.2.$^{C3.1}$

C3.1. In textbooks one often finds the term "a sufficiently small test charge". In carrying out the experiment, one has to make compromises in order to have sufficient sensitivity and precision for the measurement.

> This field would pull the charged carrier towards the plate which is nearest. In order to prevent that, the carrier must be placed precisely at the center between the plates (in an unstable equilibrium).

2. *In addition*, we use the current source *I* to apply a voltage *U* between the plates *C* and *A*. This produces a new field-line pattern, as in part c of the figure. It is the result of a superposition of the patterns b and a (cf. Fig. 3.9, below). The carrier is now pulled upwards by the field from the condenser plates.

3. We measure the force F using the balance. Furthermore, we measure the charge Q on the carrier. This is done with the calibrated ballistic galvanometer (carrier α is contacted by the wire 2!). Finally, we measure the voltage U and the spacing l between the condenser plates.

4. From each set of four corresponding quantities (the charge Q, the force F, the voltage U and the spacing l), we compute the product $F\,l$ and the product UQ and find experimentally, within the error limits, that they are proportional to each other:

$$F\,l \sim Q\,U\,.$$

If we again set the constant of proportionality equal to 1, we obtain

$$F\,l = Q\,U\,. \tag{3.1}$$

(The *direction* of the force will be discussed in the following section.)

In this equation, the term on the left is a quantity of work; therefore, the term on the right must also be 'work'. That is, one can measure work not only mechanically as the product of force times distance, but also electrically, as charge times voltage; or, written as an equation,

$$W = Q\,U\,. \tag{3.2}$$

The charge is the time integral of a current; when the current I is constant over time, we can write $Q = I\,t$ (Eq. 2.1). Inserting this quantity into Eq. (3.1) yields

$$F\,l = UIt\,, \tag{3.3}$$

an equation which we have already encountered in Sect. 1.12.

In the experimental setup which we have used (Fig. 3.1), the electric field was homogeneous. Its field strength is given by $E = U/l$. Inserting E into Eq. (3.1) gives

$$F = Q\,E \tag{3.4}$$

(e.g. F in newton, Q in ampere second, E in volt/meter).[C3.2]

In words: The force observed is proportional to the carrier charge Q and also to the field strength E of the condenser field (which is assumed not to be modified by the *small charge* on the carrier (Fig. 3.2, part a). E is not for example the field strength of the actual field present during the measurements (part c)![C3.3]

Equations (3.1) to (3.4) are often used. We set the proportionality factor in Eq. (3.1) equal to 1. This means that the three quantities 'work', 'charge', and 'voltage' cannot be measured independently of one another; instead, work and charge are used to measure the voltage as a *derived* quantity. This makes it possible to use the same energy units for mechanical and electrical measurements (see Sect. 1.12).

C3.2. The complete equation in vector form will be given in the following section. Of course, Eq. (3.4) also holds when other units are used; it is an equation relating physical quantities, as always.

C3.3. This is not due only to the compromise mentioned in Comment C3.1; rather, the "test charge" also produces an electric field, whose magnitude is in fact rather large near the carrier ($\sim 1/r^2$, Eq. 2.15). However, this field cannot exert a force on the carrier (test charge) itself.

3.3 The General Definition of the Electric Field $\boldsymbol{E}$

The essential qualitative property of electric fields are the *forces* which they exert on electric charges *at rest*. These forces lead to the preferred directions which are so graphically apparent in the patterns of electric field lines, and they require that an electric field $\boldsymbol{E}$ be represented as a *vector*. The vector nature of the electric field $\boldsymbol{E}$ is already apparent in its defining equation. To show this, we first have to agree upon a measurement procedure for defining the charge Q (Sect. 2.11), and then we can define the field $\boldsymbol{E}$ as a vector quantity using the fundamental equation (3.4),

$$\boldsymbol{F} = Q\boldsymbol{E} \quad \text{or} \quad \boldsymbol{E} = \frac{\boldsymbol{F}}{Q}. \qquad (3.5)$$

In this relation, Q refers to a small charge on a *test object*, which does not affect the form of the field already present. Equation (3.5) contains the definition of the direction of the electric field vector: For a positive charge, $\boldsymbol{E}$ and $\boldsymbol{F}$ are parallel and point in the same direction.[C3.4]

If forces $\boldsymbol{F} = Q\boldsymbol{E}$ move a charge Q in an arbitrary field, e.g. an inhomogeneous field, along a path $\boldsymbol{s}$, then they perform the work

$$\int \boldsymbol{F}\cdot \mathrm{d}\boldsymbol{s} = Q\int \boldsymbol{E}\cdot \mathrm{d}\boldsymbol{s}\,, \qquad (3.6)$$

and from this, the voltage follows:

$$U = \int \boldsymbol{E}\cdot \mathrm{d}\boldsymbol{s} = \frac{1}{Q}\int \boldsymbol{F}\cdot \mathrm{d}\boldsymbol{s} \qquad (3.7)$$

as a derived quantity with the unit

$$1\ \text{volt (V)} = \frac{1\,\text{N m}}{1\,\text{A s}}.$$

Theoretical treatments need only describe the results of experiments, but not to demonstrate them quantitatively. As a result, in mechanics, they can use the equation $\boldsymbol{F} = m\boldsymbol{a}$ as a defining equation and place it at the beginning of the treatment, or also, in electrodynamics, they can follow the path sketched out in this section. However, whoever wants to derive the fundamental empirical facts quantitatively from experiments will have to resign themselves to following a longer and more tedious path, making use of the currently-available technical resources.

[C3.4] C3.4.This is also found experimentally from the fundamental experiment in Fig. 3.1: The "test charge" was negative, the field directed from above to below, and the resulting force was directed upwards. This agrees with Eq. (3.5), since a force opposite to the field direction acts on *negative* charges.

3.4 First Applications of the Equation F = QE

The application of Eq. (3.5) is in general not at all simple. Usually, the carrier distorts the initially-present electric field due to influence, even when it is not itself charged. The field takes on a complicated form. In such cases, one has to compute the field strength at each surface element of the uncharged carrier; then, after charging the carrier, one must multiply the charge on the surface element by that field and sum over all such products. In reality, the situation is still more complicated, since the charge on the carrier will shift the charge distribution on the electrodes that produce the field by influence (imagine a point charge in front of an uncharged metal plate: It induces a *mirror charge* of opposite sign due to influence). Integration of the MAXWELL equation (2.8) is in general possible only with the aid of complex computer programs. This tedious procedure can be avoided only in a few limiting cases; we offer two examples:

1. The *forces between two small spheres at a large distance R from each other*. A sphere carrying the charge Q has, by itself, a radially-symmetric field (compare Fig. 2.48). At a distance R, it produces an electric field of strength

$$E_\mathrm{R} = \frac{Q}{4\pi\varepsilon_0 R^2}\,. \tag{2.15}$$

After bringing up the second sphere with its charge Q', the field becomes quite different. For the special case that the two charges are equal, $Q = Q'$, it can be seen in Fig. 3.3 (where the charges have opposite signs), and in Fig. 3.4 (for charges of the same sign).

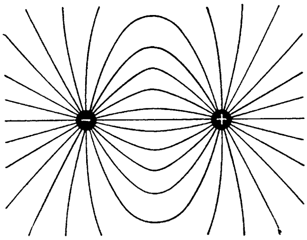


Figure 3.3 Field lines between charges with opposite signs. As in Fig. 3.4, also, they were generated by vector addition of the fields of the individual charges (the direction of the field lines is by convention from + to −, or beginning on +, ending on − charges).

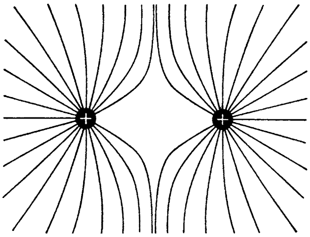
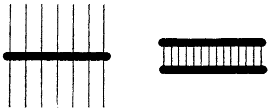


Figure 3.4 Field lines between charges of the same sign. The corresponding negative charges can be thought of as residing on the distant walls of the room.

Figure 3.5 The attraction of two condenser plates. The increased density of the field lines in the right-hand image reflects the increased field strength.



In order to apply the equation

$$\boldsymbol{F} = Q'\boldsymbol{E}\,, \tag{3.5}$$

one has to assume the original, undistorted field of the first sphere alone (Eq. (2.15)), i.e. to combine Eqns. (2.15) and (3.5). In this way, we obtain the magnitude of the force[C3.5]

$$F = \frac{1}{4\pi\varepsilon_0}\frac{Q\,Q'}{R^2}\,, \tag{3.8}$$

C3.5. COULOMB's law (Eq. (3.8)) in vector form is $\boldsymbol{F} = \dfrac{1}{4\pi\varepsilon_0}\dfrac{Q\,Q'}{R^2}\cdot\dfrac{\boldsymbol{R}}{R}$, where $\boldsymbol{F}$ is the force which Q exerts on Q', and $\boldsymbol{R}/R$ is the unit vector which points from Q towards Q'. In this formulation, in addition to the magnitude of the force, its direction and sign are also specified.

and for its direction, the result that it points along the line connecting the two charges (spheres). For charges of the same sign, it is a repulsive force, and for opposite signs, an attractive force. This law was first stated in the form $F = \pm Q_s Q_s'/R^2$ by COULOMB.[C2.10] It was the culmination in 1785 of a century of experimental research. Nevertheless, most treatments of electromagnetism put it at the very beginning.

2. The *attraction of the two plates (electrodes) in a flat parallel-plate condenser*. A plate (with charge Q) by itself produces the field sketched at the left in Fig. 3.5. The field lines can be thought of as ending on distant charges of opposite sign on the walls of the room or other distant surfaces. Compare Fig. 2.10. The field is still homogeneous at not-too-large distances in front of and behind the plate. There, its magnitude is

$$E = \frac{D}{\varepsilon_0} = \frac{1}{\varepsilon_0}\frac{Q}{2A}\,. \tag{3.9}$$

This field is to be used in applying Eq. (3.4). It acts on the charge Q on the second plate with the force[C3.6]

$$F = Q\frac{1}{\varepsilon_0}\frac{Q}{2A} = \frac{1}{2\varepsilon_0}\frac{Q^2}{A}\,. \tag{3.10}$$

C3.6. The two charges have opposite signs, which gives rise to the attractive force. For simplicity, POHL considers only the magnitudes here.

The charges on the second plate modify the field drastically (Fig. 3.5, right side).[C3.7] All the field lines on the upper side of the plate now vanish. What remains is the well-known homogeneous field of a flat parallel-plate condenser.

C3.7. Since the charges are free to move over the surfaces of the plates, they no longer remain on both faces, but instead are now concentrated on the inner surfaces, each facing the other plate. This however has no influence on the new field distribution, as one can readily understand by considering insulating plates with fixed charges instead.

Now we change the meaning of the symbol E. We again use it to denote the field of the complete condenser. Then we find

$$Q = \varepsilon_0 E A\,, \tag{2.4}$$

$$F = \frac{1}{2}QE = \frac{\varepsilon_0}{2}E^2 A \tag{3.11}$$

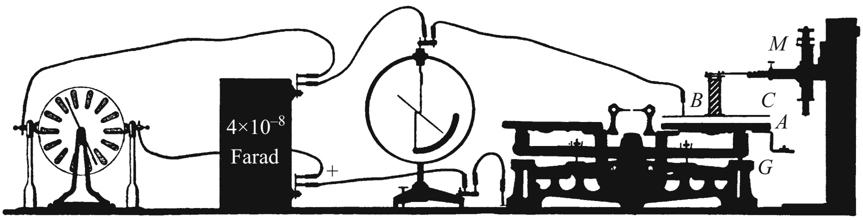


Figure 3.6 The mutual attraction of two condenser plates C and A. B is an insulating post, M is a screw micrometer with a mm scale and a vernier, G is a weight. *Numerical example*: $A = 20 \times 20\,\text{cm}^2 = 4 \cdot 10^{-2}\,\text{m}^2$; plate spacing $l = 10.2\,\text{mm} = 10.2 \cdot 10^{-3}\,\text{m}$; voltage $U = 7500\,\text{V}$[C3.8]

$$F = \frac{8.86 \cdot 10^{-12}}{2} \cdot \frac{\text{A s}}{\text{V m}} \cdot \frac{5.63 \cdot 10^{7}\,\text{V}^2 \cdot 4 \cdot 10^{-2}\,\text{m}^2}{1.04 \cdot 10^{-4}\,\text{m}^2}$$
$$= 9.6 \cdot 10^{-2}\,\frac{\text{W s}}{\text{m}} = 9.6 \cdot 10^{-2}\,\text{N}\,.$$

When the electric force becomes greater than the weight of G, the plate A begins to move upwards.

C3.8. Instead of the influence machine and the large condenser which is needed to smooth its voltage output, one could of course use a high-voltage power supply. Note the robust kitchen scale which permits a sensitive measurement of the forces.

or

$$F = \frac{\varepsilon_0}{2}\,\frac{U^2 A}{l^2}\,, \tag{3.12}$$

i.e. the force is proportional to the square of the voltage U and inversely proportional to the square of the spacing l of the plates.

Figure 3.6 shows a setup for testing this equation. It is intended in particular to give a feeling for the orders of magnitude involved. For precision measurements, one must use a flat parallel-plate condenser with a “corona ring” here also (Fig. 2.44).

According to Eq. (3.12), the force increases in inverse proportion to the square of the spacing of the condenser plates. Therefore, to produce large forces, condensers with very small spacings have been constructed. A highly-conducting plate is set onto a poor conductor, both with smooth surfaces. Figure 3.7 shows a metal plate M in contact with a lithography stone St. Both have a surface area of around $20\,\text{cm}^2$. The stone’s weight is about 2 N (mass $m = 200\,\text{g}$). When a voltage of 220 V is applied, the stone “sticks” to the metal plate. It can be lifted together with the metal plate using the handle. Of course, this condenser is not insulated. In our

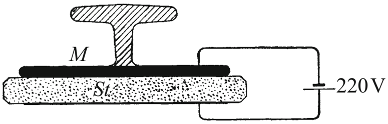


Figure 3.7 The mutual attraction of two condenser plates which are made of a good conductor M and a poor conductor St. As a result of the unavoidable roughnesses on their surfaces, the spacing is very small at some points, and the electric fields are therefore very large there **(Video 3.1)** **(Exercise 3.1)**.

Video 3.1:
“Forces in an Electric Field”
http://tiny.cc/xaggoy
Note the protective resistors! Note also that the experiment in Fig. 3.7 is reversed: the stone is lifted and pulls the metal plate below it upwards.

example, a current of several 10^{-6} A flows across it. The human body only feels currents of more than 3 to 5 mA (Fig. 1.29). We could readily use it as a "connecting wire" instead of the metal wires shown in Fig. 3.7, and thus make the stone "stick" onto the metal plate.

3.5 Pressure on the Surfaces of Charged Bodies. Reduction of Their Surface Tension

Pressure is generally defined by the ratio

$$p = \frac{\text{Force } F \text{ which acts perpendicular to a surface}}{\text{Surface area } A} \,.$$

For the homogeneous field of a parallel-plate condenser, we thus find from Eq. (3.11)

$$p_{\mathrm{e}} = \frac{\varepsilon_0}{2} E^2 \,. \tag{3.13}$$

Here, E is the field strength directly on the inner surfaces of the plates.

We want to apply this equation to the case of a charged sphere of radius r. The voltage between the sphere and the distant carriers of the opposite charge is U. Then on its surface, the field strength is

$$E = \frac{U}{r} \,. \tag{2.20}$$

We insert this value into Eq. (3.13) and obtain for the *pressure on the surface of the charged sphere*:

$$p_{\mathrm{e}} = \frac{\varepsilon_0}{2} \frac{U^2}{r^2} \,. \tag{3.14}$$

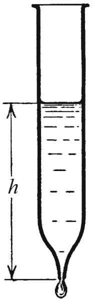
This pressure is directed outwards[1], it acts like a *reduction of the surface tension* ζ. The surface tension by itself produces a pressure that is directed inwards, $p_0 = 2\zeta/r$ (see Vol. 1, Sect. 9.5). In the presence of an electric field, the remaining inwardly-directed pressure is reduced to

$$p = \frac{2\zeta}{r} - \frac{\varepsilon_0}{2} \frac{U^2}{r^2} \,. \tag{3.15}$$

The reduction of the surface tension by an electric field can be demonstrated in many ways, e.g. using the setup shown in Fig. 3.8. The spray nozzle of a glass container emits a jet, which becomes a stream of droplets as the depth h of the water decreases. The fact that the water coalesces into droplets is a result of its surface tension. Now, we produce an electric field between the water and the walls of the room using an influence machine. Immediately, the water again flows out of the nozzle as a smooth jet.

[1] This is a convenient but lax way of putting it. The pressure is not 'directed', but rather the associated force.

Figure 3.8 The influence of an electric field on the surface tension of water (GEORGE MATHIAS BOSE, 1745)



3.6 GUERICKE's Levitation Experiment (1672). The Elementary Charge, $e = 1.602 \cdot 10^{-19}$ A s.

A particularly important application of the equation $\boldsymbol{F} = Q\boldsymbol{E}$ for physics is the "levitation experiment". It is the original form of the arrangement shown in Fig. 3.1. A light-weight charge carrier is brought into a vertically-directed electric field. Suppose that the carrier is negatively charged and the condenser electrode above it is positive. Then its weight F_{G} pulls the carrier downwards and the force

$$F = QE \quad \text{or} \quad F = Q\frac{U}{l} \tag{3.1}$$

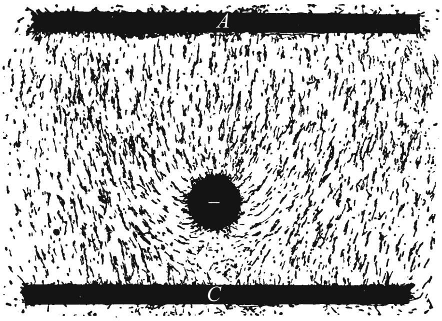
pulls it upwards (compare the field lines in Fig. 3.9). In the limiting case

$$F_{\mathrm{G}} = Q\frac{U}{l}, \tag{3.16}$$

there will be an "equilibrium", and the carrier is "levitated". Then we can compute the charge Q on the carrier from its weight F_{G} and the field strength U/l.

For demonstration experiments, light-weight objects which would fall only slowly in air, such as feathers or cotton, gold leaf flakes,

Figure 3.9 The electric field lines (lines of force) in the levitation experiment. One can practically "see" how the charge carrier is pulled upwards, although the field which enters in Eq. (3.16) is the homogeneous field of the empty condenser.



Video 3.2:
"Soap Bubbles in an Electric Field"
http://tiny.cc/9aggoy

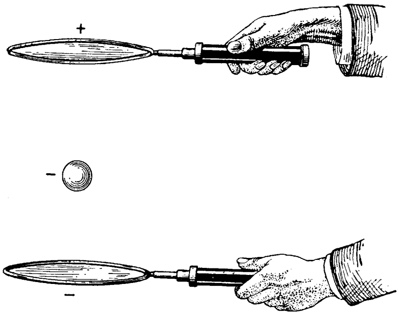


Figure 3.10 A charged soap bubble levitated in an electric field **(Video 3.2)**

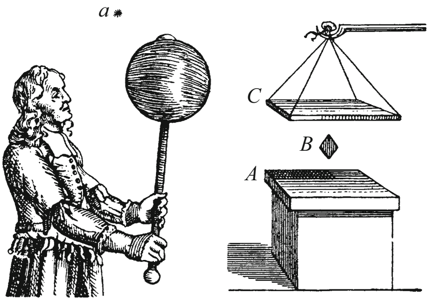


Figure 3.11 Old representations of the levitation experiment: at the right, by Benjamin Wilson (1746); at the left by Otto von Guericke (1672) (*B*: gold-leaf flakes; *a*: downy feathers). *"Plumula potest per totum conclave portari"*

soap bubbles etc. are most suitable. These carriers are charged and then caught in the electric field between two charged condenser plates (Fig. 3.10). The electric field strength is varied by changing the spacing of the plates. (The field as shown in Fig. 3.10 is not homogeneous; otherwise its field strength would be independent of the spacing of the plates.) In this way, we can cause the carriers to rise, to sink, or to remain levitated (i.e. to "float" at a fixed height). To simplify the setup, the upper plate in Fig. 3.10 is often left off; then the ceiling of the room takes on its function. In this form, the levitation experiment was first demonstrated by Otto von Guericke[2], (1602–1686), in the year 1672 (Fig. 3.11).

The levitation experiment can be readily repeated on a greatly reduced scale; instead of the soap bubble in Fig. 3.10, one can use small liquid droplets, usually oil or mercury, with diameters of less than 1 µm. They are charged by contact to a solid body (by "static electricity"). This can be accomplished by letting them be carried in an air jet through the nozzle of a spray bottle; friction with its walls separates the necessary charge. The condenser plates have a spacing of ca. 1 cm. The motions of the charged droplets in the electric field are observed through a microscope. The weight of the droplets can be obtained by a microscopic measurement of their diameters[3]. The volume can be computed from the diameter and yields the weight F_{G}

[2] Guericke was mayor of Magdeburg, 1646–1681. See also Vol. 1, Sect. 9.9, and J. L. Heilbron, p. 216, cited in the footnote in Sect. 2.17.

[3] Usually, however, their diameters are found from their sinking speed in air (Vol. 1, next-to-last paragraph of Sect. 10.3).

after multiplication by the density of the liquid and the acceleration of gravity. A very fundamental result emerges from such experiments with small, but still readily visible charge carriers (R. A. MILLIKAN, 1910; he continued the classic experiments of J. S. E. TOWNSEND, 1897, and J. J. THOMSON, 1898):

An object can accept or release electric charges only as integral multiples of the 'elementary charge' $e = 1.602 \cdot 10^{-19}$ A s. In spite of numerous efforts, no one has ever been able to observe a smaller charge than $1.602 \cdot 10^{-19}$ A s on a positively- or negatively-charged object. For this reason, we call the *charge* $e = 1.602 \cdot 10^{-19}$ A s the *elementary electric charge*. It is the smallest individually *observed* negative or positive electric charge. For example, an electron possesses exactly one negative elementary charge.

The 'Millikan experiment' is not difficult to carry out. It should be found in every beginning physics teaching laboratory. It is most impressive when observed individually through the microscope. With a projection microscope, even gentle air currents in the condenser disturb the observations. They arise from the intense light source which is required for projection.[C3.9]

C3.9. Today, this disturbing effect can be easily avoided by combining a television camera with the microscope.

The *cathode-ray tube* (or oscilloscope) already mentioned in Sect. 1.11 can be conveniently used to demonstrate the force on individual electrons in an electric field (Eq. 3.5). The schematic is shown in Fig. 3.12. The electrons (of charge e and mass m) which emerge from the hot cathode C pass through a tube F and through a hole in the anode A. This tube and the anode together serve as an electric lens and form an image of the small opening in the anode on the fluorescent screen S. Along their paths from the anode A to the screen S, the electrons pass through two parallel-plate condensers whose field directions are rotated by 90° relative to each other. We show in the figure only one of these, namely DD'. Using its electric field, the electrons can be deflected within the plane of the paper (the other condenser produces deflections perpendicular to the plane of the paper). In this manner, one can combine the two perpendicular deflections ("x and y axes"). The deflections are proportional to the field strengths E produced by the voltages between the condenser plates. For the deflection x of the path (Fig. 3.13), we find

$$x = \frac{1}{2}\frac{e}{m}E\frac{y^2}{u^2}\,. \tag{3.17}$$

Derivation: Let the electron pass along the length y of the condenser at a velocity u in the time $t = y/u$. During this time, the electron falls through a distance $x = \frac{1}{2}a\,t^2$. Here, a is the acceleration of the electron in the direction opposite to the electric field $\boldsymbol{E}$ which produces it. From the fundamental equation of mechanics, the force $F = m\,a$ and thus with Eq. (3.4), $e\,E = m\,a$. Inserting the values of a and t yields Eq. (3.17). (Compare this also with the parabolic flight path treated in Vol. 1, Sect. 4.9.)

Finally, we make one more not-unimportant remark: A *dropper bottle* can dispense its medicine only in "elementary quanta", namely

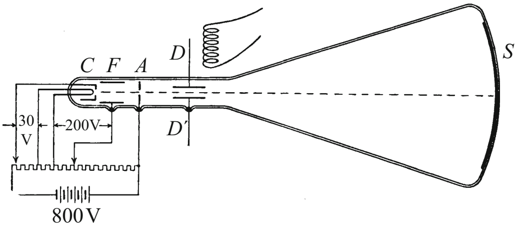


Figure 3.12 A cathode-ray tube with a heated cathode. The latter consists of a glowing tungsten filament C just behind a negatively-charged collimator (focussing) tube F ("WEHNELT cylinder"). In modern versions, the cathode itself is not imaged on the fluorescent screen, but instead a strongly demagnified image of the cathode. The condenser plates D and D' ('deflection plates') serve to deflect the electron beam.

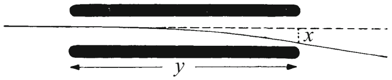


Figure 3.13 The deflection of electrically-charged beams by the homogeneous electric field of a parallel-plate condenser

as individual droplets. This of course does not prompt us to assume that droplets exist independently inside the bottle. Similarly, the Millikan levitation experiment doubtless shows that there is a lower limit to the divisibility of electric charges. But it by no means proves the same subdivision of charge within the interiors (or on the surfaces) of objects! The existence of individual, countable charge "quanta" within a charge carrier remains for the moment only a very useful assumption.[C3.10]

[C3.10] C3.10. This assumption appears to have been confirmed only recently by the following observation: The electrical properties like currents or capacitances of metallic or semiconducting particles of the order of 100 nm in diameter can be varied over orders of magnitude by changing their electric charges (the "COULOMB blockade" effect, "single-electron transistors"; see M. A. Kastner, *Physics Today*, Jan. 1993, p. 24).

3.7 The Energy of the Electric Field

In an empty space of volume V, suppose there is an electric field of magnitude E. How much energy is contained in this field?

We imagine that the field is produced by a charged parallel-plate condenser. Let the area of its plates be A and their spacing l, so that the field volume is $V = A\,l$. The one plate attracts the other and pulls it along a distance $\Delta\, l$, performing work in the process, for example work of lifting according to the schematic of Fig. 3.14. It does this with the constant force

$$F = \frac{\varepsilon_0}{2} E^2 A\,, \tag{3.11}$$

since the charge Q and thus the field strength $E = D/\varepsilon_0$ remain unchanged. For the work performed, equal to the energy previously

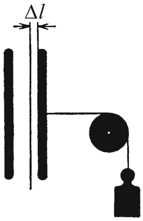


Figure 3.14 The derivation of the energy of an electric field. The plates are *not* connected to a current source.

stored in the electric field, we then find

$$\Delta W_\mathrm{e} = F\,\Delta l = \frac{\varepsilon_0}{2} E^2 A\,\Delta l\,,$$

or, integrating ($\Delta l \to \mathrm{d}l$) over the full spacing of the plates from 0 to l,

$$W_\mathrm{e} = \frac{\varepsilon_0}{2} E^2 V \tag{3.18}$$

$$\left(\varepsilon_0 = 8.86 \cdot 10^{-12}\,\mathrm{A\,s/V\,m}\right).$$

Equation (3.18) holds quite generally, in spite of our having derived it for a specific case.C3.11

Only small amounts of energy can be stored in practice in the form of electric fields. For example, in one liter ($= 10^{-3}\,\mathrm{m}^3$) at a technically practicable field strength of $E = 10^7\,\mathrm{V/m}$, the stored energy is only $0.44\,\mathrm{W\,s}$.

Equation (3.18) for the energy of an electric field is often written differently, i.e. by making use of Eqns. (2.4) and (2.2):

$$W_\mathrm{e} = \frac{1}{2} Q U\,, \tag{3.19}$$

and continuing with the aid of Eq. (2.10),

$$W_\mathrm{e} = \frac{1}{2} C U^2\,. \tag{3.20}$$

Here, Q denotes the charge on a condenser of arbitrary form, U its voltage and C its capacitance.C3.12

3.8 The Electric Potential. Equipotential Surfaces

To represent electric fields, in addition to the field-line patterns, effective use is often made of the electric “equipotential surfaces” (contour maps). In Fig. 3.15, we show an electric field between a plate and a wire which is parallel to it. Directly above the plate, there is a small

C3.11. This means that it holds also for inhomogeneous electric fields. In a volume element $\mathrm{d}V$, containing an electric field E, the energy

$$\mathrm{d}W = \frac{1}{2}\varepsilon_0 E^2 \mathrm{d}V$$

is stored. Often, one quotes instead the *energy density*:

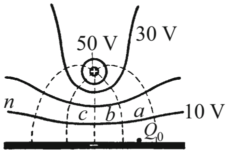
$$\frac{\mathrm{d}W}{\mathrm{d}V} = \frac{1}{2}\varepsilon_0 E^2\,.$$

From this, it follows that in general

$$W_\mathrm{e} = \frac{\varepsilon_0}{2}\int\limits_V E^2 \mathrm{d}V\,.$$

C3.12. The fact that the two equations (3.19) and (3.20) hold not only for empty parallel-plate condensers, but in general, i.,e. for example when the condenser contains a dielectric, can be derived from the following considerations: The energy W_e stored in a charged condenser is completely converted into JOULE heat by discharging the condenser through a resistor R. If we insert Eq. (2.23) for I into the expression for the power, $\dot{W}_\mathrm{e} = I^2 R$ (Eq. (1.7)) and integrate over time, we obtain these two equations.

Figure 3.15 A schematic drawing of electrical equipotential surfaces



charge carrier with the charge Q_0 (test charge). This carrier can be moved to the point a; that would require performing the work W. In electric units, this work is Q_0U (Eq. (3.2)), where U is the voltage between the starting and end points of the path along which the carrier is moved. We then repeat the same experiment but with different starting points above the surface of the plate and different end points in other regions of the field. Each time, we stop when the work $W = Q_0U$ has been performed. The carrier is then at one of the end points $a, b, c, \dots n$. The set of all such points which can be reached by performing the same amount of work is called an *equipotential surface*.

To characterize an equipotential surface, we make use of the ratio

$$\frac{\text{Work } QU \text{ performed } \textit{against} \text{ the force } QE \text{ of the field}}{\text{Charge } Q \text{ on the carrier}} = U\,. \tag{3.21}$$

U is the voltage between the equipotential surface and the *conventional reference point*; in Fig. 3.15, this is the plate. This electric tension or *voltage* is also called the *potential difference* and is denoted by the symbol $\Delta\varphi$. The reference point is often connected to the earth's surface ("grounded"); then the potential at a particular point in the field means the voltage between that point and the earth's surface. The *"potential" is thus a name for the voltage between an arbitrary point within a field and a conventional reference point, and the equipotential surfaces are surfaces of constant potential*. In Fig. 3.15, and in other fields produced by two charged bodies, the positive sign of the potential implies a negative charge at the reference point.

Justification: If the carrier in Fig. 3.15 has a positive charge Q_0, then the work $W = Q_0 \cdot \varphi$ must be performed *on it* in order to bring it from a negative reference point to the equipotential surface. Then W is positive. Therefore, in Eq. (3.21), the numerator and the denominator have the same sign and the voltage $U = \Delta\varphi = W/Q_0$ is positive. (At the same time, the path integral $\int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s}$ (in Eq. 2.3) is negative, since the direction of the field points by convention from + to − and therefore the carrier is moved *against* the field into regions of increasing potential; i.e. $\boldsymbol{E}$ and $\mathrm{d}\boldsymbol{s}$ point in opposite directions.)

The *potential* φ can be specified at only a *single* point within the field; this is not true of the *voltage* U, which always applies between *two* points (it is the potential *difference*). If one refers to a voltage as a potential, this implies that a reference point was previously agreed upon (often taken without specific mention to be the earth's surface). Unfortunately, the words 'potential' and 'voltage' are often used loosely as synonyms and not correctly distinguished. The *potential difference* between two points within a field is the *voltage* between those two points.[C3.13]

Example: Consider the potential distribution (potential field) of a charged metal sphere (carrying a charge Q_0 at radius r). Assume that the reference point (where the potential is *zero*) is at $R = \infty$. The electric field strength within the sphere is zero, and outside the sphere, it is given by $E = Q_0/4\pi\varepsilon_0 R^2$ for $R \geq r$ (Eq. (2.15)). Then we find in the region outside the sphere the potential:

$$\varphi = -\int_{\infty}^{R} E\,\mathrm{d}R = -\frac{Q_0}{4\pi\varepsilon_0}\int_{\infty}^{R}\frac{1}{R^2}\mathrm{d}R = \frac{1}{4\pi\varepsilon_0}\frac{Q_0}{R}\,, \qquad (3.22)$$

and within the sphere,

$$\varphi = \frac{1}{4\pi\varepsilon_0}\frac{Q_0}{r} = \text{const}\,. \qquad (3.23)$$

Thus, for $Q_0 > 0$, the potential decreases as $1/R$ outside the sphere and is constant in its interior. The equipotential surfaces are concentric spherical shells (for $R \geq r$).

C3.13. The relationship between the voltage U or potential difference $\Delta\varphi$ (between two points 1 and 2), and the path integral $\int \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{s}$, is thus found (in agreement with the above justification in petit type) to be

$$U = \Delta\varphi = -\int_{1}^{2}\boldsymbol{E}\cdot\mathrm{d}\boldsymbol{s}\,.$$

In Eq. (2.3), for simplicity, the sign was not taken into account.

3.9 The Electric Dipole. Electric Dipole Moments

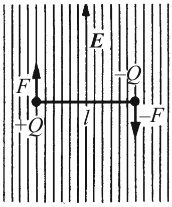
The fundamental equation $\boldsymbol{F} = Q\boldsymbol{E}$ for the appearance of forces in an electric field requires not only the field itself, but also a body carrying an electric *charge*. Considered superficially, this would appear to contradict our long-established experience: We observe forces on light-weight *uncharged* bodies in electric fields. Imagine some *scraps of paper* in the neighborhood of a piece of amber which has been rubbed with fur; or the *dancing dolls* below a glass plate which has been charged with static electricity.

To understand these phenomena, we need two *new concepts*: the *electric dipole* and the *electric dipole moment*. We suppose that in Fig. 3.16, two "pointlike" electric charge carriers with the charges $+Q$ and $-Q$ are attached to the ends of an extremely thin and ideally insulating rod of length l. This dumbbell-shaped object is called an *electric dipole*. Its electric field is similar to those shown in Figs. 2.8 and 3.3.

We further imagine this dipole to be oriented as in Fig. 3.16 with its long axis (dipole axis) perpendicular to the field lines of a homogeneous electric field. Then a torque acts on it:

$$M_{\text{mech}} = 2QE\frac{l}{2} = QlE\,. \qquad (3.24)$$

Figure 3.16 An electric dipole which is oriented perpendicular to electric field lines



We call the product $Q\,l$ the *electric dipole moment* p of the dipole (unit: ampere second meter, A s m). The electric dipole moment must be described as a *vector*; its *direction* is defined to be *along the connecting line between the two charges from* $-$ *to* $+$. This then leads in general to

$$\boldsymbol{M}_{\text{mech}} = \boldsymbol{p} \times \boldsymbol{E}\,. \tag{3.25}$$

C3.14. For the vector product, see Vol. 1, Comment C6.1.

This vector product[C3.14] means that the mechanical torque which acts on the electric dipole is maximal when $\boldsymbol{p}$ is perpendicular to $\boldsymbol{E}$ (as drawn in Fig. 3.16 and assumed in Eq. (3.24)), and that it vanishes when $\boldsymbol{p}$ is parallel to $\boldsymbol{E}$.

The idealized dipole as defined above is not found in reality. However, positive and negative charges of equal magnitude can be localized at a particular separation within matter in a number of ways, and a measurement procedure can be used to define the dipole moment of a structure of this type. Such a procedure makes reference to an experiment in mechanics:

In Fig. 3.17, a rod S is mounted at the end of a spoke R. The force couple $\boldsymbol{F}$ and $-\boldsymbol{F}$ produces a torque $\boldsymbol{S} \times \boldsymbol{F}$ where $\boldsymbol{S}$ is the vector that points from the point of action of $-\boldsymbol{F}$ to the point of action of $\boldsymbol{F}$. The length of the spoke R is then completely unimportant.

Now we suppose that in some arbitrary body, N dipoles have been formed by some sort of localization of charges. Each of them experiences a torque $\boldsymbol{M}'_{\text{mech}}$ in an applied electric field $\boldsymbol{E}$. All of these individual torques can be added vectorially, in spite of their different distances from the mutual axis of rotation of the body. We thus obtain the overall (observable) torque:

$$\boldsymbol{M}_{\text{mech}} = \sum \boldsymbol{M}'_{\text{mech}} = \sum (\boldsymbol{p}' \times \boldsymbol{E})\,, \tag{3.26}$$

or

$$\boldsymbol{M}_{\text{mech}} = \boldsymbol{p} \times \boldsymbol{E}\,. \tag{3.27}$$

Figure 3.17 For a torque, only the length of the lever arm S is important, not the length of the spoke R

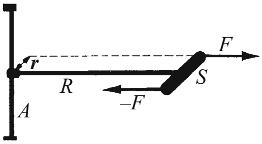
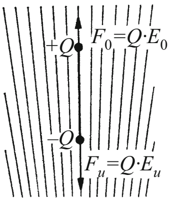


Figure 3.18 An electric dipole in an inhomogeneous electric field (the field direction is upwards (+*x* direction), and *E* decreases with increasing *x*)



Here, ***p*** denotes the total, macroscopically observable electric dipole moment of the body which is the vector sum of the individual, unknown microscopic dipole moments within it.

This overall dipole moment can always be represented in terms of an idealized dumbbell-shaped dipole: Two pointlike charges $+Q$ and $-Q$ are fixed at a separation l. The rod in this dumbbell points along the direction of its electric dipole moment.

This defining equation suggests a measurement procedure for electric dipole moments (which is however unimportant in practice): We mount an object on a rotatable axis perpendicular to an electric field and find its rest position. Then we rotate its axis by 90° out of the rest position and measure the required torque as the product of force and force arm (lever). This torque is then divided by the strength of the homogeneous electric field E to give the overall dipole moment p.

So much for an electric dipole or a body with an electric dipole moment in a *homogeneous* field. The field acts on the dipole, producing a torque, and orients the dipole with its axis parallel to the field direction, as long as its rotation is not hindered. The same holds true in an *inhomogeneous* electric field. The dipole in Fig. 3.18 is supposed to have already oriented itself along the field direction (positive x direction). In addition, however, in an inhomogeneous field, something new occurs: In an inhomogeneous field, along the direction of increasing field $\partial E/\partial x$, there is a *force*

$$F = p\,\frac{\partial E}{\partial x}\,. \tag{3.28}$$

Derivation: The upper + charge experiences the force QE_o upwards, while the lower − charge experiences the force $-QE_\mathrm{u}$, downwards. Combining these, the overall force

$$F = Q(E_\mathrm{o} - E_\mathrm{u}) \tag{3.29}$$

acts on the dipole. Furthermore, we have

$$E_\mathrm{o} = E_\mathrm{u} + \frac{\partial E}{\partial x}\,l\,. \tag{3.30}$$

Equations (3.29) and (3.30) together yield Eq. (3.28).

3.10 Induced and Permanent Electric Dipole Moments. Pyro- and Piezoelectric Crystals

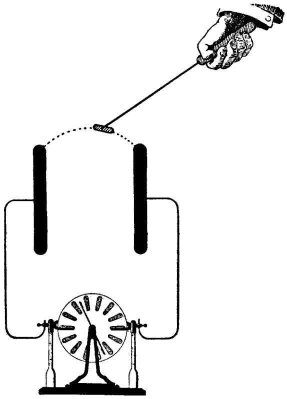
We have thus far introduced the concepts of the electric dipole and its dipole moment without referring to experiments. Now we ask the question: How can we produce electric dipoles within matter in practice? We must distinguish between two cases:

1. *Electric dipole moments produced by influence*. Every material body acquires an electric dipole moment in an external electric field, due to *influence*: The field displaces the positive and negative charges relative to each other within every material body which is brought into it. In a conductor, they move to its *surfaces*; in an insulator, there are displacements *within* the individual *molecules*, giving rise to a *polarization* (or *"electrification"*) of the dielectric (Fig. 2.56).

As a result of this "induced" dipole moment, produced by influence, oblong bodies orient themselves in *all* electric fields parallel to the field direction[4]. The parallel orientation is energetically favored relative to all other possible orientations (Fig. 3.19). This is how e.g. *field-line patterns* are made visible with fibrous powder. In *inhomogeneous* fields, all bodies are in addition pulled towards regions of higher electric field strength, regardless of their shapes.

At the delimiting boundaries of the field (e.g. on the plates of a condenser), well-conducting bodies immediately become charged, and then, as "charge carriers", they fly over to the opposite electrode, where the process repeats itself with opposite sign. In the case of insulators or poor conductors, the charging requires some time (several seconds). During this time, the object sticks to the electrode. This can be seen very clearly in a shadow projection using small cotton balls.

Figure 3.19 Forces on an uncharged object in an electric field; top, a small, oblong object made of metal, or an insulator which is mounted on a rotation axis perpendicular to the plane of the page (a "versorium": WILLIAM GILBERT, 1600, physician in London and originator of the word "electric").[C3.15]



C3.15. See A. Loos, "*Vier Jahrhunderte Spannung*", *Physik in unserer Zeit* **31**, 159 (2000); English: See https://en.m.wikipedia.org/wiki/William_Gilbert_(astronomer) .

[4] The polarization of the oblong body attains its maximum value in this orientation, and with it, the induced dipole moment (this is because the *depolarization factor* (see Sect. 13.6) has its smallest value along the long axis).

2. *Permanent electric dipole moments*. a). In every charged *condenser*, charges of both signs are displaced relative to one another. As a result, most charged condensers have their own electric dipole moment. It is lacking only when one of the condenser electrodes surrounds the other in the form of a closed hollow box.

Unfortunately, an electric field produces a dipole moment even in every *uncharged* condenser due to *influence*. Therefore, in Sect. 3.9, we did not start from experiments; every one of our dipoles would have still moved in the field even if they had originally not possessed permanent dipole moments.

b). We put a mixture of liquid wax and resin into an electric field and allow it to solidify there. Its induced electric dipole moment (due to influence) is *frozen in* and thus has become permanent. As a result, the solidified object (preferably cut into the form of long sticks after solidification) acts as an *electret*. It appears to be a good electric insulator with positive electric charges at one end and negative charges at its other end. These charges can be measured in an influence experiment using a ballistic galvanometer (Sect. 1.10). The leads of the galvanometer are connected to metal cartridges and these are wiped simultaneously over the two ends of the electret; the *induced* charges separated in the cartridges by influence are then measured by the galvanometer.

Electrets of this type maintain their polarization for years, as long as they are kept in a well-fitted metal protective housing; otherwise, over the course of time, they attract charge carriers (ions) from the air and thereby cover their ends with a layer of charges of opposite sign, so that their electric dipole moments are no longer detectable to the outside world.

c). *Pyroelectric* crystals, e.g. tourmaline, have permanent electric dipole moments owing to the arrangement of their charged microscopic components (F. U. T. Aepinus, 1756). Their directions fall along a polar crystalline axis; in the case of an oblong tourmaline crystal, for example, along its long axis. Normally, this permanent dipole moment is not apparent, due to the covering layer of ions as mentioned above. It becomes active only when one changes the elementary electric dipoles by *thermal expansion or contraction*. For example, if we dip an oblong tourmaline crystal of ca. 5 cm length into liquid nitrogen, then only part of the covering layer compensates the dipolar charges. The crystal now appears to be a good electret, attracting scraps of paper, etc. (Fig. 3.20).

Liquid nitrogen is often clouded by fine ice crystals. They can be removed by dipping a tourmaline electret into the liquid.

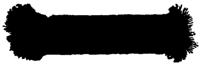
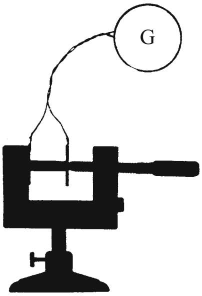


Figure 3.20 A coating of powder attracted to the ends of an electret (tourmaline)

Figure 3.21 A tourmaline crystal acquires an electric dipole moment when pressed (detected by a ballistic galvanometer or a static voltmeter)



Pyroelectric crystals are at the same time *piezoelectric*, i.e. they change their electric dipole moments also as a result of *mechanical deformation* (compression or tension). This can be demonstrated by again using a rod-shaped tourmaline crystal. One fixes it between two insulated electrodes, connects them to a galvanometer, and compresses the crystal along its long axis using a vice (Fig. 3.21).[C3.16]

[C3.16] Conversely, piezoelectric crystals are deformed by an electric field (they become longer or shorter). We can for example produce sound waves (ultrasound generator), or control very small length changes by an applied voltage, as in a scanning tunnel microscope.

Exercises

3.1 Refer to Video 3.1: In the video, the brass plate is lifted by applying a voltage of 460 V, which is just sufficient to make it stick to the stone when lifted. Compute the average spacing l between the stone and the brass plate. For simplicity, we assume that the entire voltage drop occurs across the gap l. We want to check this assumption. The brass plate has a volume of $5 \cdot 8 \cdot 0.5\,\text{cm}^3$ and a density of $\varrho = 8\,\text{g/cm}^3$ (Sect. 3.4). (Note that the roles of the metal plate and the stone slab are reversed in Fig. 3.7 as compared to the video!)

3.2 A water droplet is charged with a single electron. How large must the radius R of the droplet be if it can just be levitated by the electric field of the earth (130 V/m, Sect. 2.12)? (Sect. 3.6).

3.3 A parallel-plate condenser with a capacitance C is connected to a current source with the voltage U. Find the change in the energy W_e stored in the condenser when the spacing of its electrodes is changed by the factor $1/n$ (Sect. 3.7).

For Sect. 3.7, see also Exercises 2.11, 2.15, and 2.16;
for Sect. 3.8 see also Exercise 2.7.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_3) contains supplementary material, which is available to authorized users.

4 The Magnetic Field

4.1 The Production of Magnetic Fields by Electric Currents

(H. Ch. OERSTED, 1820). The introductory summary in Chap. 1 mentioned three characteristics of electric currents in a conductor: 1. A magnetic field; 2. a heating effect; and 3. chemical changes in the conducting medium.

These three characteristics are by no means of equal importance. Chemical changes are lacking in the technically most important conductors, the metals. Heating of the conductor (JOULE heating) may also be lacking under certain conditions (superconductivity$^{C4.1}$). However, the *magnetic field* is present in all cases. *A magnetic field is the inseparable companion of every electric current.*

C4.1. For remarks on superconductivity, see Comment C10.3.

A magnetic field – just like an electric field – can exist in empty space (the *vacuum*). The presence of air molecules (cf. Fig. 2.14) is practically unimportant. Magnetic fields, like electric fields, can be studied only experimentally. We observe different phenomena in a magnetic field from those in ordinary (field-free) space. This is the decisive fact here. The most important of these phenomena which we have already encountered was the orientation of iron filings or powder into chain-like patterns, providing images of the magnetic field lines.$^{C4.2}$

C4.2. For the concept of *field lines*, see Comment 2.1.

We now want to delve deeper into the nature of magnetic fields. We start by considering some of the typical geometric forms taken by magnetic fields:

The magnetic field lines of a long, straight current-carrying conducting wire are concentric rings (Fig. 1.4).

If the conductor is formed into a ring, we obtain field lines as shown in Fig. 4.1. The "circles" appear to be pushed eccentrically outwards and to be somewhat deformed. We wind the wire around a circular form, so that several circular windings are adjacent to each other (Fig. 4.2). Now, the field-line patterns from each winding are superposed to give an overall field pattern. One could imagine that each circular winding has its own current source; more conveniently, they are connected in series so that the same current flows through all of them, forming a *coil* (by winding the wire continuously onto a spool in a helical pattern; cf. Figs. 4.3 and 4.4).

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_4

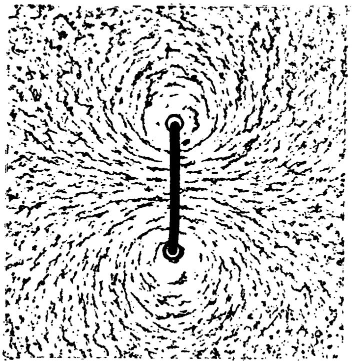


Figure 4.1 The magnetic field lines from a current-carrying circular ring, made visible with iron filings

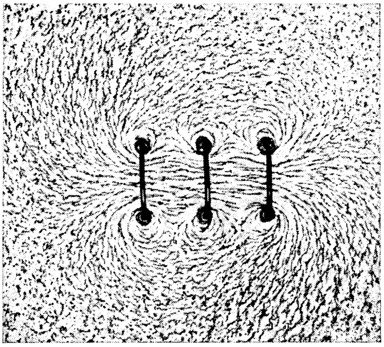


Figure 4.2 The magnetic field lines from three parallel circular rings, each carrying the same current

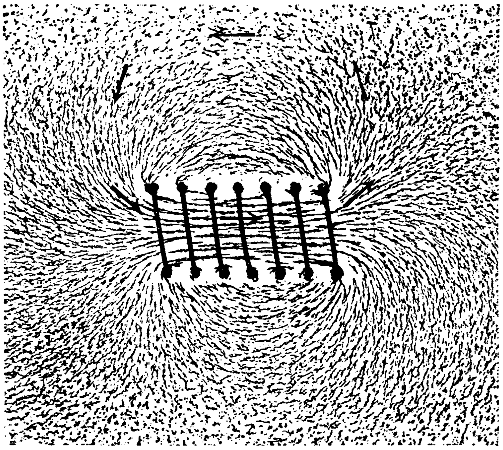


Figure 4.3 The magnetic field lines from a short current-carrying coil. The arrows denote compass needles, and their arrowheads are the north poles. Here, the + pole of the current source is connected to the end of the coil at the upper left.[C4.3]

One end of a compass needle normally points to the north; this end is called its north pole and is often marked by an arrowhead. In the magnetic field of a coil, the compass needle points along the direction

[C4.3] C4.3. If we look at the coil from the left, the current is seen to be circling in a clockwise sense around its axis. Within the coil, it produces a magnetic field which points to the right in the plane of the page. The left-hand end of the coil is its south pole (just like the left-hand end of the compass needle within the coil).

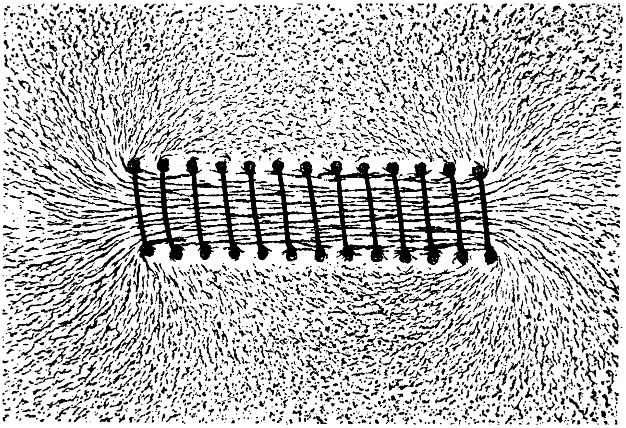


Figure 4.4 The magnetic field lines from a long current-carrying coil. In the interior of the coil, there is a (nearly) homogeneous magnetic field.[C4.4]

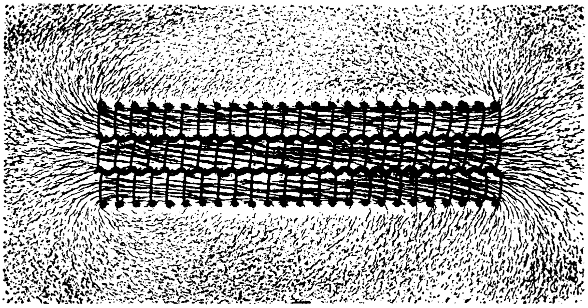


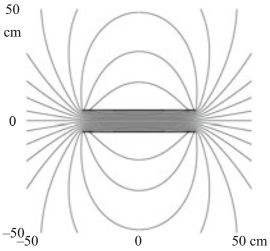
Figure 4.5 The magnetic field of a bundle of long, thin coils. The individual coils were completely separate in this model experiment. The apparent zig-zag shaped connections are only simulated by the accumulation of iron filings between neighboring wires.

of the field lines (Fig. 4.3). *The direction in which the arrowhead points is defined by convention to be the positive direction of the field*.

The same field as with a single large coil can be obtained with a bundle of identical long, thin coils, if the bundle has the same cross-sectional area and all the coils are carrying the same current. Figure 4.5 shows a field-line pattern obtained in this way. Compare it to Fig. 4.4. This experimental finding is readily understood: In Fig. 4.6, we have drawn a bundle of coils in cross section; for simplicity, the cross sections of all the individual coils have been chosen to be square. In the inner parts of this drawing, one can see that neighboring currents are opposite and thus cancel. Only the thick black windings around the perimeter of the bundle contribute to the field. Only this current path need be considered.

At the ends of the coils, field lines project out into the open space outside the coils. They emerge not only through the two open ends of

C4.4. The field distributions shown here, obtained using iron filings to trace the pattern of the field lines, are quite similar to the real field distributions, as can be seen here in the figure:



This figure shows the field distribution of a coil as computed from the MAXWELL equations (Sect. 6.5); the ratio of length to diameter of the coil was taken to be 5 : 1, corresponding to the coil shown in Fig. 4.4 (computed by J. A. Crittenden, Cornell University). The inhomogeneous regions of the field (near the poles) are discussed in more detail in Sect. 8.6.

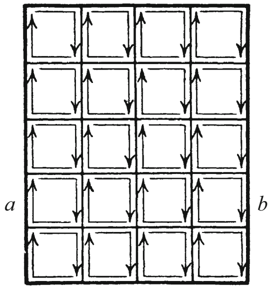


Figure 4.6 Schematic of a bundle of *long*, square coils. a–b corresponds to the plane of the picture in Fig. 4.5.

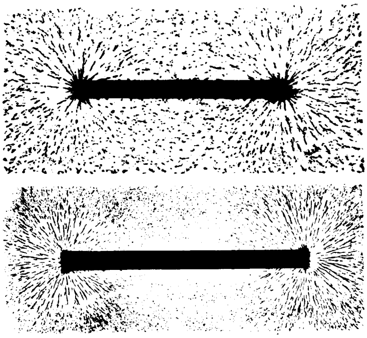


Figure 4.7 *Top*: Accumulation of iron filings around a current-carrying coil; *Bottom*: The poles of a permanent bar magnet made visible with iron filings (the magnet is made of a ceramic: ferrite powder with a binder, shaped and sintered in a magnetic field). This magnet has the same geometry as the coil in the top image. The bar is composed of 100 flat slabs which have been pressed together.

the coils, but also near the ends, out of the sides of the coils through their windings. *These regions where the field lines emerge from a coil are called its* poles, in analogy to a permanent bar magnet. A current-carrying coil behaves very much like a bar magnet: If it is suspended or balanced on an axis horizontally, the coil will orient itself like a compass needle in the north-south direction. If iron filings are scattered over it, a coil will attract them to its poles (Fig. 4.7). The middle section outside a coil will remain free of iron filings. The field lines emerge only at the regions called "poles". As a coil becomes longer and longer, the poles become less and less important relative to the field in the interior of the coil. Compare for example Figs. 4.3 and 4.4.

It is also possible to fabricate *coils which have no poles*. The coil must be wound in the form of a closed ring. Figure 4.8 shows an example. In the coil shown, the cross section of the windings is constant all around the ring, but this is not necessary; by the right choice of the spacing of neighboring windings, it is possible to produce coils without poles but with a variable cross-sectional area.

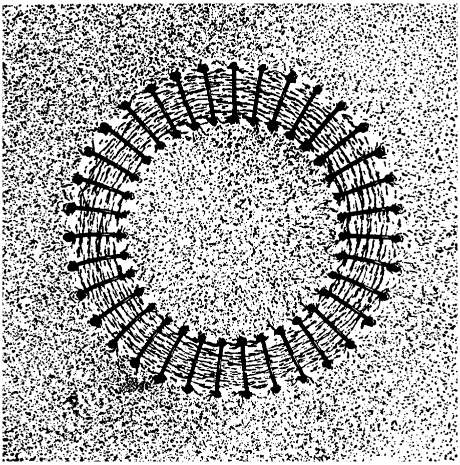


Figure 4.8 Magnetic field lines in the field of a current-carrying ring coil (torus)$^{C4.5}$

C4.5. Such "toroidal" magnetic fields are used for example to contain the hot plasma for research into nuclear fusion (Tokamak principle). For these experiments, large superconducting coils with diameters of several meters are employed.

We summarize: *The geometric distribution of the magnetic fields of current-carrying conductors is determined solely by the geometry of the conductors themselves.*

In the case of *long, thin* coils, the magnetic field lines within the coil are practically straight lines, apart from their polar regions. We are then dealing with a *homogeneous field*.$^{C4.6}$ The homogeneous magnetic field of a long coil plays a similar role in the treatment of magnetic fields as does the homogeneous electric field of a flat parallel-plate condenser in the discussion of electric fields.

C4.6. The field becomes more and more homogeneous the longer the coil (see the figure in Comment C4.4.)

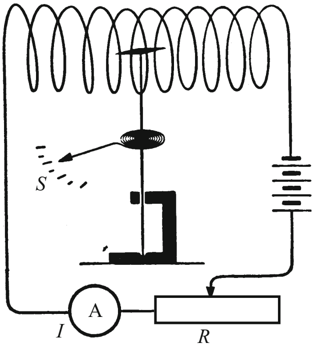
The magnetic fields of *bar magnets* or, in general, *permanent* magnets are in no way different from those of current-carrying coils; that is, we can replace the magnetic field of any bar magnet in the space surrounding it by the field of a coil of similar size and geometry. We need only adjust the distribution of the windings correctly. We will discover the reason for this similarity in Sect. 4.4.

In Chap. 2, we made use of the *homogeneous* electric field of a parallel-plate condenser to define two quantities which allowed us to quantitatively characterize any electric field; those were the field quantities $\boldsymbol{E}$ and $\boldsymbol{D}$. Analogously, we will now use the *homogeneous* magnetic field of a long coil to define two new quantities with which we will be able to quantitatively characterize any magnetic field. These are the fields $\boldsymbol{H}$ and $\boldsymbol{B}$. Initially, we will consider only the field $\boldsymbol{H}$.

4.2 The Magnetic Field $\boldsymbol{H}$

Like the electric field, the magnetic field must be described quantitatively by a *vector*. This follows from the readily-seen *preferred directions* of the magnetic field lines. The vector is called the *magnetic field* $\boldsymbol{H}$.

Figure 4.9 The quantitative determination of the magnetic field of a long coil (schematic!). (R is the variable resistor or rheostat used to adjust the current I through the field coil)



The *magnitude of an electric field, i.e. its electric field strength* can be measured in volt/meter (V/m). Correspondingly, the *magnitude of a magnetic field, called the magnetic field strength*, can be measured in ampere/meter (A/m). This can be confirmed by a new series of experiments. They require two components, namely

1. long coils of various designs; and

2. an arbitrary indicator for the magnetic field.

The indicator need only be able to identify two spatially or temporally separated magnetic fields as *equal*; it is not required to measure them, but only to determine that the two fields have the same strength.

As our indicator, we choose a small magnetic needle[1] (compass needle) connected to the axle of a spiral-spring torsion balance (Fig. 4.9). The resting position of the needle is determined by the relaxed position of the spiral spring. (For simplicity, we neglect the influence of the earth's magnetic field.)

We insert this magnetic needle into the homogeneous field of a coil and rotate the pointer until it is in its resting position on the scale S (spring relaxed). Then we *wind up* the spring by rotating the pointer until the needle is oriented perpendicular to the field lines (and thus perpendicular to its own resting position). The required *tension* of the spring can then be read off the scale; it is a measure of the torque required to rotate the needle to a position perpendicular to the magnetic field .[C4.7]

C4.7. This is completely analogous to the torque that an electric dipole experiences in an electric field (cf. Fig. 3.16 and Eq. (3.25)).

We now replace the coil by a series of different ones, and continue repeating the experiment through the whole series of coils. These may have different cross-sectional areas A, different lengths l and different numbers of turns N. Some of these coils have a single layer of windings, while some have many layers. By varying the current through the coils (using the rheostat R), we adjust the torque required to rotate the compass needle to a perpendicular orientation so that it

[1] The needle must be small compared to the linear dimensions of the field, in order to have sufficient spatial resolution.

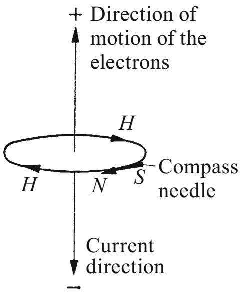


Figure 4.10 The determination of the direction of the magnetic field in the neighborhood of an electron current (see also Fig. 1.5)

has the same value as with the first coil in every case. This equality of torques means that the *magnetic fields* also have equal strengths.

In this way, we find experimentally a very simple result: The magnetic fields are *equal*, as long as the quantity

$$\frac{\text{Current}\, I \times \text{Number of turns}\, N}{\text{Coil length}\, l}$$

has the same value. The cross-sectional area and the number of winding layers have no influence on the field strength. The *homogeneous magnetic field of a long coil is determined uniquely by the ratio NI/l; or, in words, by the "current times number of turns divided by the length of the coil"*.[C4.8] For this reason, we use the ratio NI/l in order to arrive at a first definition of the magnetic field strength in a long coil; it is:

$$H = \frac{NI}{l}\,. \tag{4.1}$$

(The product NI of the current and the number of turns in the coil winding is sometimes abbreviated as 'ampere-turns'). The unit of H is thus 1 A/m. The direction of the magnetic field vector is parallel to the long axis of the coil. Its sign is indicated in Fig. 4.3 (see also Fig. 4.10).

The next step takes us to an important generalization. By making a comparison to the homogeneous field of a long coil, we can measure the field $\boldsymbol{H}$, its magnitude and direction, at any arbitrary location: We replace its individual small regions (which are practically homogeneous) by the equally strong and similarly oriented field of a small, long coil and determine the vector $\boldsymbol{H}$ for this coil using Eq. (4.1) and the direction of the coil and its current.

As a measurement prescription, one can proceed in several different ways: We could for example calibrate the arrangement in Fig. 4.9 and thus convert it into a *magnetometer*. The calibration is accomplished by varying the current I in a long coil and thus changing its field strength H as defined by Eq. (4.1). We find that the torque required to orient the needle is proportional to H. For example, the demonstration model used in the lecture hall at Göttingen has a calibration factor of 50 A/m per angular degree.

C4.8. One must take care to avoid the end regions of the coils in these measurements; there, the field strength decreases markedly (see the figure in Comment C4.4).

C4.9. By comparison with Fig. 4.3, we can see that the geographic north pole of the earth is in fact a magnetic south pole!

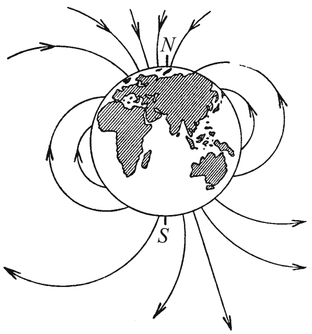


Figure 4.11 The magnetic field lines of the earth's field. The arrows indicate the direction of the field.[C4.9]

With a magnetometer which has been calibrated in this way, we want to measure the magnetic field of a bar magnet at a point P about 10 cm in front of its north pole. We bring the needle with its spring relaxed along the direction of the field lines, i.e. in the direction of the field. Then we turn the needle by rotating the frame until the needle is perpendicular to the field and read off the scale S that the spring has been tensioned by rotating through an angle of 10 angular degrees. Therefore, at the point P, the strength of the field is $H = 500$ A/m.

Analogously, we can measure the earth's magnetic field. Figure 4.11 shows a schematic drawing of its field lines. The component which is parallel to the surface of the earth is called the horizontal component; in Göttingen, it has a value of about 16 A/m.

Magnetometric measurements are time-consuming and therefore somewhat tedious. They are however indispensable when one needs to measure very small fields; they are then carried out with a different technique, as will be discussed in Sect. 8.6. In the great majority of cases, one *calculates* the field strength. Examples can be found in Sect. 6.3.

4.3 The Motion of Electric Charges Produces a Magnetic Field. ROWLAND's Experiment (1878)

An electric current within a conductor consists of moving charges along its length (Sect. 2.10). Now we consider a surprising fact: It is simply this *motion* of the electric charges which produces the magnetic field. No other details of the process are important. The conductor, e.g. a copper wire, acts only as a guide for the moving charges; or, roughly speaking, as a pipe through which they flow. This is shown by ROWLAND's experiment:

Figure 4.12 shows a plan view of a *circular carrier of electric charge* at the outer edge of an insulating disk (which is shaded in the figure). This conducting ring is interrupted by a narrow slit between *a*

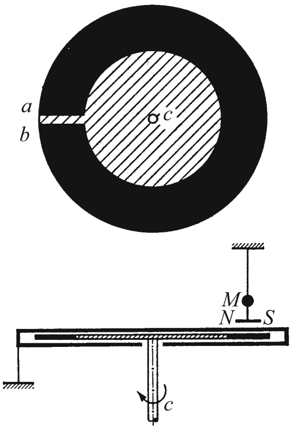


Figure 4.12 ROWLAND's experiment (the diameter of the ring-shaped charge carrier is ≈ 20 cm)

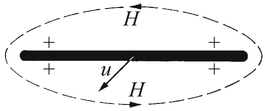


Figure 4.13 The magnetic field lines which surround a positive charge that is moving at a velocity u towards the observer

and b. The lower image shows the same charge carrier from the side; it is mounted on a vertical shaft and enclosed in a grounded *metal housing*. Between the ring-shaped carrier and the housing, a voltage of about 10^3 V can be applied, so that the ring then carries a positive charge Q of around 10^{-7} A s. At the point M, there is a sensitive magnetometer, indicated schematically in the figure; it could be a compass needle with a mirror and light pointer. (In its resting position, the long axis of the needle is perpendicular to the plane of the page). The charged carrier rotates at N revolutions within the time t; its rotational frequency is $N/t \approx 50$ Hz. During the rotation of the charged ring, the needle indicates the presence of a magnetic field.[C4.10] Thus, *a charge set mechanically in motion produces a magnetic field, just like a charge moving through a conductor*. Its magnetic field lines are drawn schematically in Fig. 4.13. They surround the cross-sectional area of a section of the rotating ring (enlarged in the drawing), to the right of the axle c; the section is moving towards the observer with the velocity u.

C4.10. The magnetic field produced in this apparatus was 50 000 times weaker than the horizontal component of the earth's magnetic field (see Fig. 4.11) at the location of the experiment (≈ 16 A/m in Berlin; see H. A. Rowland, *Am. J. of Science and Arts*, 3rd Series, Vol. 15, p. 30 (1878)). More sensitive measuring instruments such as are available today, e.g. a HALL probe or a SQUID (superconducting quantum interference device), were not available back then!

In the second part of the experiment, the charge carrier is discharged; then leads are attached at the points a and b in Fig. 4.12 and a current I is passed through the ring (≈ 10^{-5} A). It produces an identical magnetic field to that of the rotating charged carrier. We see that

$$I = Q\frac{N}{t}\,. \tag{4.2}$$

We then introduce the velocity u of the carrier and its path $l = 2\pi r$ into this equation, obtaining

$$u = \frac{Nl}{t} \quad \text{or} \quad \frac{N}{t} = \frac{u}{l}\,. \tag{4.3}$$

Inserting this quotient into Eq. (4.2) yields the important relation:[C4.11]

$$I = Q\frac{u}{l}\,, \tag{4.4}$$

or in words: *A charge Q which is moved along the path l at the velocity u acts as a current $I = Qu/l$.* We will return to this significant experiment in Chap. 7.

C4.11. This is the definition of an electric current:

$$\frac{Q}{t} = I = \frac{Q\,l}{l\,t} = \frac{Q}{l}u\,.$$

4.4 The Magnetic Fields of Permanent Magnets are also Produced by the Motion of Electric Charges

In our first experiments, we produced magnetic fields with the aid of electric currents in metal conductors (wires). Then we saw that magnetic fields also result from the mechanically-produced motion of charges. Now, we consider as the third possibility the oldest method of obtaining magnetic fields: their production by permanent magnets. How do the magnetic fields of permanent magnets come about?

We again refer to experiment and take a current-carrying coil (Fig. 4.14). Its magnetic field is supposed to be just detectable at the point *P*; a compass needle set up there shows a small angular deviation α. How could we amplify the magnetic field and increase the signal α?

Either: We increase the current *I* through the coil, or its number of turns *N* or both. In any case, we would increase the product *NI*, the number of 'ampere-turns' of the coil.

Or: We introduce a previously non-magnetic piece of soft iron into the coil, an 'iron core'.

This leads us to conclude that *the iron effectively increases the number of ampere-turns*. Of course, it neither increases the number of turns in the coil winding, nor does it increase the current, which we monitor continuously with an ammeter. Therefore, within the iron, there must be microscopic currents flowing along invisible paths in the same sense as the macroscopic current in the coil. Their 'ampere-turns' add to those of the windings of the coil. This notion causes no difficulties: According to ROWLAND's experiment, we need only assume that some sort of orbital motions of electric charges are present

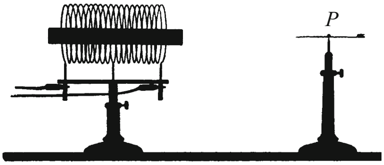


Figure 4.14 Inserting an iron core acts like an increase in the number of turns or the current (i.e. the 'ampere-turns') in a coil (solenoid)

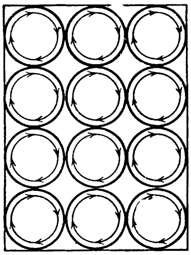


Figure 4.15 A rough schematic of ordered molecular currents

within the iron. We know that electrons are present in all material objects. We can imagine their motion within the iron to be microscopic circular (orbital) motions .[C4.12] This preliminary but already quite serviceable concept is called the model of *molecular currents*. It can be represented roughly by a drawing as in Fig. 4.15. Compare this figure with the cross-section through a bundle of thin coils in Fig. 4.6.

C4.12. This idea occurred to Ampère as early as 1826. He called them 'circuital currents' and presumed that such currents were responsible for "material magnetism", i.e. the magnetism of permanent magnets.

These molecular currents must have been present in every piece of iron even before it is put into a magnetic field; but they are on the average not ordered. Only in the magnetic field of the coil do they become ordered: Their rotational axes all line up parallel to the long axis of the coil. The individual molecular-current orbits act like the small rotatable coil in Fig. 1.10.

We can reduce the magnetic field of the coil either by pulling the iron core back out or by interrupting the current through the coil. Then the field produced by the iron for the most part disappears, but not completely. The majority of the molecular currents rotate back to their original disordered arrangement; only a certain portion maintain the orientation that they received in the magnetic field. The iron is then said to exhibit "remanent" magnetism (Chap. 14). It has become a *permanent magnet* (like a compass needle).

Only one single point is essential in this model: The existence of some sort of orbital motions of electric charges (e.g. the electrons) in the interior of the iron. This decisive point can be tested experimentally: The mechanical angular momentum of the orbiting charges can be demonstrated and measured (Einstein-de Haas experiment, 1915).

We remind the reader of the following experiment in mechanics: A person is sitting on a swivel chair and is holding a rotating object, e.g. a wheel. Its plane of rotation is in some arbitrary orientation relative to the figure axis, and the chair is at rest. Then the person moves the rotational axis of the wheel so that it is perpendicular to the figure axis (Fig. 4.16). This tipping of the rotating wheel gives the person an *angular momentum*, causing a rotation (of person and chair) around the figure axis. This rotation gradually comes to a stop due to friction in the bearings of the chair and with the air.

Now we imagine that the person on a swivel chair is replaced by an iron rod, and the wheel by the disordered orbiting elementary charges. The iron rod is hanging as shown in Fig. 4.17 along the

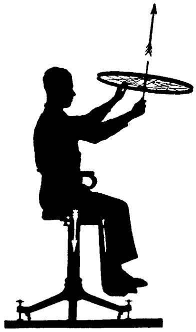


Figure 4.16 Conservation of angular momentum

long axis of a coil. When the current in the coil is switched on, the rotational planes of all the orbiting charges are oriented perpendicular to the axis of the rod and the coil. The iron rod takes on the resulting angular momentum and rotates (just like the person in the swivel chair). Details of the experimental results and analysis are given in Sect. 14.9.

For the practical implementation of the experiment, we could use a current pulse lasting only around 10^{-3} s (discharging a condenser through the coil). This method makes make use of the small fraction of the molecular currents that remain parallel after the current goes to zero, and thus give rise to the remanent magnetism of the iron. If the current were kept constant, the unavoidable inhomogeneities of the magnetic field in the coil would disturb the measurement. The iron rod would be gradually pulled into the region where the magnetic field is strongest, as in the experiment illustrated in Fig. 1.11 (solenoid action).

For their fundamental experiment, Einstein and de Haas[C4.13] applied a sinusoidally-varying current to the coil, and the corresponding sinusoidal magnetic field $H(t)$ acted as a driving force for the weakly-damped torsional pendulum formed by the iron rod and its suspension fiber (cf. Fig. 4.17). At its resonance frequency (see Vol. 1, Sect. 11.10 and this volume, Sect. 11.7), the resonance enhancement (Vol. 1, Eq. (11.7)) makes the oscillation amplitude of the torsional pendulum large and readily detectable in the laboratory or lecture

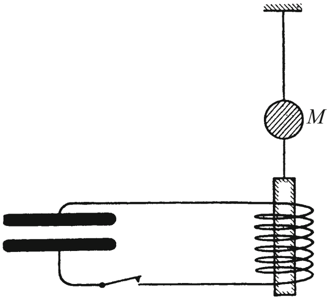


Figure 4.17 Schematic of the experiment to demonstrate the molecular currents in iron. The homogeneity of the magnetic field in the coil is not sufficient to observe the rotation using continuously flowing currents; therefore, the field-reversal or the pulsed-field method must be used. (M is a mirror for the optical pointer)

C4.13. In the original Einstein-de Haas experiment, the resonant alternating-current method was used. For the history and description of these experiments, see V.Ya. Frenkel, Usp. Fiz.Nauk 128 (1979), p. 545 (in English; online at http://www.physics.umd.edu/grt/taj/411c/EinsteindeHaas.pdf). In modern demonstration experiments of the EINSTEIN-DE HAAS effect (the “magneto-mechanical parallelism”), this method is also employed. In this way, the experiment can be readily demonstrated in the lecture room.

room. From this amplitude (corrected for the resonance enhancement, see Eq. (3) in the paper by Frenkel, Comment C4.13.), the angular momentum can be obtained, and comparison to the maximum magnetization M_0 of the iron rod (measured separately) yields the "gyromagnetic ratio", i.e. the ratio of the magnetic moment to the angular momentum of the "molecular currents". See also Sect. 14.9.

After this experimental detection of the angular momentum of the molecular currents, we can say today with certainty that the magnetic fields of permanent magnets are also produced by the motions of electric charges.

In earlier times, *magnetic substances* were held to be the origin of the magnetic fields of permanent magnets. Similarly to the electric field lines, it was presumed that the magnetic field lines would begin on one body and end on another. At their ends, there would be magnetic "charges" or poles of opposite signs (called "magnetic monopoles"). All efforts to separate the magnetic charges have been futile to this day. Even the relatively primitive model of molecular currents makes this failure understandable. In this model, a permanent magnet is in the end the same as a bundle of current-carrying thin coils, and there, we know that there are only closed field lines in the form of loops without beginning or end. A refinement of this model will be discussed in Chap. 14.

Induction Phenomena

5

5.1 Preliminary Remark

For an observer at rest, electric charges at rest produce only an electric field, but charges in motion produce a magnetic field in addition. This relation between magnetic and electric fields follows from the ROWLAND experiment. A still closer connection between the two kinds of fields is revealed by *induction phenomena*. This chapter presents the experimental facts relating to induction in a vacuum (i.e. – for all practical purposes – in air). Chapters 6 and 7 will then deal with their evaluation and explanation.

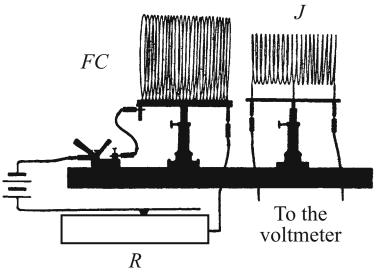
5.2 Induction Phenomena (M. FARADAY, 1832)

We start with an inhomogeneous magnetic field of arbitrary origin, e.g. the field from the short current-carrying coil (*the field coil FC*) in Fig. 5.1. A second coil *J* is placed in this magnetic field; it will be called the *induction coil* from now on. Its ends are connected to a voltmeter with a short response time. With this setup, we carry out a series of experiments which can be organized into three groups:

1. We leave the induction coil at rest in the magnetic field and change the strength of the field by varying the current through the field coil (rheostat *R* and switch).

2. We change the position of the induction coil and the field coil relative to one another by sliding or rotating one of them.

Figure 5.1 Induction experiments



© Springer International Publishing AG 2018
K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_5

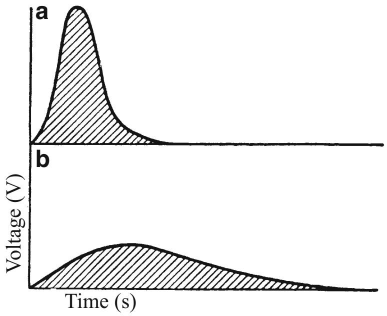


Figure 5.2 Two *voltage impulses* of the same magnitude $\int U\,\mathrm{d}t$, measured in volt second (V s)

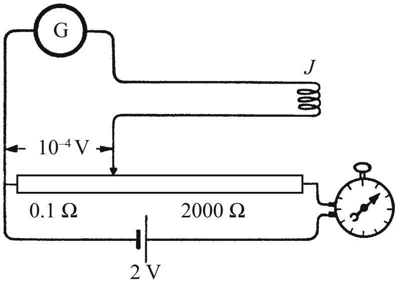


Figure 5.3 A rotating-coil galvanometer *G* connected to an induction coil *J* (cf. Fig. 5.5) is calibrated in volt second for the measurement of voltage impulses

3. Instead of the induction coil *J* as drawn, we use a ring-shaped coil made of flexible insulated wire. We deform this ring-shaped induction coil in the magnetic field, i.e. we vary its cross-sectional area by moving some parts of its windings relative to other parts.

In all three cases, we observe an *induced voltage* between the ends of the windings of the induction coil *J during* the action. Its magnitude depends on the speed of the process. For example, when one coil is rotated rapidly, we observe a voltage curve as shown in Fig. 5.2a: high voltages during a short time period. When the motion is slower, we see a curve like that shown in Fig. 5.2b: lower voltages over a longer time.

The area of the shaded regions is the time integral of the voltage ($\int U\,\mathrm{d}t$), and is also called the *voltage impulse*; it is measured e.g. in volt second (V s). This quantity is analogous to the time integral of the current, measured for example in ampere second (A s), which we treated in detail in Sect. 2.11.

For a quantitative investigation of induction phenomena, we measure the voltage impulse from the induction coil using a galvanometer with a slow response time (ballistic galvanometer). *In the following sections, we thus use the ballistic galvanometer in a different way from that in Chap. 2, where we measured current impulses in order to determine the quantity of electric charge.*

Its calibration in V s is carried out analogously to the calibration in A s which was described in Sect. 1.10 (cf. Fig. 5.3). During short, precisely measured time intervals, we apply known voltages to the galvanometer.

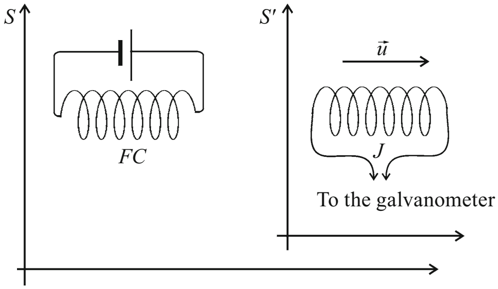


Figure 5.4 In the frame of reference S, the field coil *FC* is at rest and the induction coil *J* is moving at the velocity $\boldsymbol{u}$. In S', the situation is reversed. The galvanometer is in each case in the frame of reference in which the observer is at rest.

A known voltage of suitable magnitude is obtained from a voltage divider circuit as shown in Fig. 1.27.
We observe the galvanometer deflection α for various values of the product Ut, then take ratios B_U = (Voltage impulse Ut/Impulse deflection α), and obtain the same value in all cases, e.g.

$$B_\text{U} = 2.4 \cdot 10^{-5} \, \frac{\text{V s}}{\text{scale division}} \, .$$

This is the ballistic calibration factor of the galvanometer.

We make use of the calibrated galvanometer and repeat the three experiments described above. This leads us to an important discovery: *In the experiments of the second group, only the relative motion between the induction coil and the field coil plays a role.* In order to emphasize the significance of the statement printed in italics, we mention that it led EINSTEIN to the "principle of relativity"[1].[C5.1]

C5.1 This principle, formulated only for uniform, linear motions, along with the fact that light always propagates with the same velocity in vacuum, independently of the state of motion of the light source which emits it, together form the basis for EINSTEIN's theory of special relativity (*Annalen der Physik*, Vol. 17, (1905), p. 891). A good account of the historical development of this theory is given by F. Hund, "*Geschichte der physikalischen Begriffe*", *BI-Hochschul-taschenbücher*, Vols. 543 and 544, Mannheim, 2nd edition, 1978. English: See e.g. M. BORN, "*Einstein's Theory of Relativity*" (Dover Publications, 1965) (available online for download at https://archive.org/details/einsteinstheoryo00born); or https://en.wikipedia.org/wiki/History_of_special_relativity .

[1] The relativity principle mentioned here can be expressed in terms of the following two equivalent statements: 1. "It is impossible to determine by experiment whether one is at rest or in a state of uniform motion". 2. "When two experiments are carried out under the same conditions in two frames of reference which are moving relative to each other at a constant velocity, both experiments will lead to the same conclusions". In order to make the relativity principle clear in terms of the setup for an experiment of group 2 as shown in Fig. 5.1, the one frame of reference S is chosen so that the field coil *FC* is at rest in it. The induction coil *J* is then moving in that frame at the velocity $\boldsymbol{u}$, e.g. to the right. In order to measure the induced voltage impulse, the galvanometer is located in frame *S* (Fig. 5.4). In the other frame of reference S', the induction coil *J* is at rest and the field coil is moving with an equal but opposite velocity, e.g. to the left; then the measurement is carried out with a galvanometer which is at rest in S'. In both frames of reference, the observers measure a voltage impulse when the coils are well separated from each other. Their descriptions are different: The observer in S says that the induction coil is moving through the magnetic field; the other observer, in S', says that the magnetic field within the fixed induction coil is changing. The relativity principle now postulates, in agreement with experiment, that both observers measure an identical voltage impulse in spite of their different descriptions of the process.

As a result, we can always discuss these experiments together with those of group 1. We need only to change our frame of reference and to consider the experiments of group 2 from the point of view of the induction coil (that is, in S'; see the footnote). There the induction coil is at rest; only the magnetic field which penetrates it changes, as in the experiments of group 1.

Quantitatively, we will describe separately induction that is due to a change in the current through the field coil, or alternatively to a change in the distance between the two coils or the relative orientations of the field coil and the induction coil: 'Induction in conductors at rest' in Sect. 5.3, and 'Induction in conductors in motion' in Sect. 5.5. This separation is quite important for our understanding of the phenomena. A summary will follow in Sect. 5.6. Furthermore, experiments from the second group will be treated in detail in both frames of reference, in terms of the theory of relativity, in Chap. 7.

5.3 Induction in Conductors at Rest

In order to treat induction quantitatively, we first make use of the homogeneous magnetic field within a long field coil (solenoid). Its field strength is given by

$$H = \frac{NI}{l} \,. \tag{4.1}$$

Furthermore, we employ induction coils of differing geometries and numbers of turns N_J. The first induction coil J surrounds the field coil FC on its outside (Fig. 5.5, left), while the second is completely inside the homogeneous magnetic field in the interior of the field coil (Fig. 5.5, right). A third induction coil has the form of a flat rectangle and consists of several hundred turns of insulated wire. This third coil can either be pushed into the field coil from the side between two of its windings, so that part of its cross-section is inside the field coil, or else it can be put completely inside, with varying angles relative to the field lines, and fixed in a tilted position (it can be rotated around the axis a, cf. Fig. 5.6). In all cases, the induction coil encloses a magnetic field of cross-sectional area A, measured in the plane perpendicular to the field lines. Examples:

The induction coil of cross-sectional area A_J is supposed to be entirely inside the field coil (Fig. 5.5, right). Then $A = A_\mathrm{J}$ when the two coils are oriented with their axes parallel. A is equal to $A_\mathrm{J}/\sqrt{2}$ when the angle between the axes of the coils is 45°, and is equal to $A_\mathrm{J}/2$ when half the area of the induction coil projects into the interior of the field coil through a slit in its side.

If the induction coil surrounds the field coil on its outside (Fig. 5.5, left), then A is equal to A_FC, the cross-sectional area of the field coil, independently of its orientation, etc.

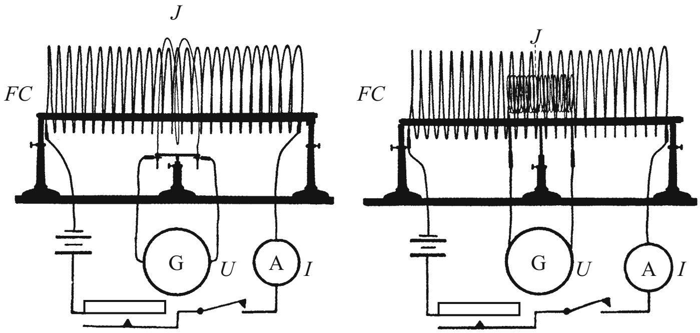


Figure 5.5 Setups for the experimental derivation of the law of induction (**Video 5.1**)

Video 5.1: "Induction in Conductors at Rest"
http://tiny.cc/ebggoy
For the experimental setup, see Fig. 5.8.

Figure 5.6 A section through a rectangular induction coil which is located within a homogeneous magnetic field and rotated relative to the field by a certain angle

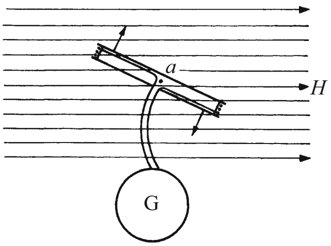
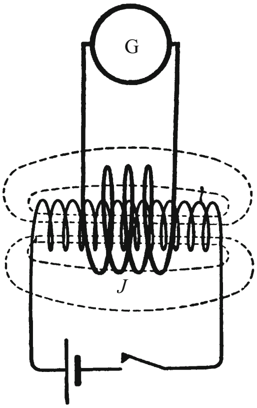


Figure 5.7 An unwanted reduction of the induced voltage impulse due to "backwards-running" (returning) field lines from a short field coil



In the latter case, one has to take care to avoid the error source illustrated in Fig. 5.7: A reduction of the induction signal due to "backwards-running" field lines (compare Fig. 4.3). The diameter of the induction coil must not exceed the diameter of the field coil by too great a margin.

Now to the experiments: The current I in the field coil is alternately switched on and off, so that its magnetic field builds up or dies out. In each case, we observe a voltage impulse $\int U\mathrm{d}t$ between the ends of the induction-coil windings. We find that it is *proportional*

1. to the current I;

2. to the ratio N/l, that is the ratio of the number of turns N to the length l of the field coil;

3. to the number of turns N_J of the induction coil J; and

4. to the cross-sectional area A of the bundle of magnetic field lines which pass through the induction coil.

All of these experimental results can be combined into a single equation. Let μ_0 be a constant factor, i.e. a constant of proportionality;[C5.2] then we have

$$\frac{\int U\mathrm{d}t}{N_\mathrm{J}A} = \mu_0 \frac{NI}{l}\,. \tag{5.1}$$

The value of the factor μ_0 in air is practically the same as in vacuum:[C5.3]

$$\mu_0 = 1.257 \cdot 10^{-6}\,\frac{\mathrm{V\,s}}{\mathrm{A\,m}}\,.$$

Common names for μ_0 are the *magnetic field constant* or the *permeability constant of vacuum*. Its ‘official’ name is the *magnetic constant*.

Equation (5.1) is one formulation of the *law of induction*.

The *sign* of the voltage impulses has been left out of our considerations so far, since it is not of decisive importance. When current through the field coil is switched on, the electric fields in the windings of the induction coil and in those of the field coil are directed *opposite* to each other; therefore, there is a minus sign on the right-hand side of Eq. (5.1). When the current in the field coil is switched off, the two fields are in the same direction. We will return to this question in Sect. 6.1 (see also Sect. 5.6).

Instead of the voltage impulse $\int U\mathrm{d}t$, in a homogeneous field, at least for a limited time, one can maintain a constant voltage U. To do this, it is only necessary to ensure that the magnetic field has a constant rate of change $\dot{H} = \mathrm{d}H/\mathrm{d}t$. Using an induction coil as in Fig. 5.5 with several hundred turns in its winding, we employ a rheostat with small steps and move its sliding contact at a constant speed; then we will obtain a constant induced voltage:

$$U = \mu_0 N_\mathrm{J} A \dot{H}\,. \tag{5.2}$$

5.4 The Definition and Measurement of the Magnetic Flux *Φ* and the Magnetic Flux Density *B*

For applications of the law of induction, we define the *magnetic flux*[C5.4]

$$\Phi = \mu_0 A\, H \tag{5.3}$$

C5.2. As we already did in Chap. 2 (Eq. (2.3)), we ignore the sign here for simplicity. This topic will be discussed in the following chapter (see also the paragraph in fine print below). The equations up to and including those of Sect. 5.5 should thus be understood as involving only *magnitudes*. This applies also to **Videos 3.2** and **5.1**, where only the *change of the sign* between switching on and switching off the field or reversing the direction of motion can be seen.

C5.3. In connection with the definition of the unit of electric current, the *ampere* (see Comment C1.6), the value of μ_0 has today been fixed by law: $\mu_0 = 4\pi \cdot 10^{-7}\,\mathrm{V\,s/A\,m}$ ($4\pi \approx 12.56637\ldots$).

C5.4. The general definition of Φ in vector notation is obtained in terms of a surface integral:

$$\Phi = \int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{A}\,.$$

and the *magnetic flux density*[C5.5]

$$\boldsymbol{B} = \mu_0 \boldsymbol{H}\,. \tag{5.4}$$

The *unit* of Φ is seen to be 1 volt second (V s), and the unit of B is thus 1 V s/m^2 = 1 tesla (T). (Sometimes, the older cgs unit Gauss (G) (1 Gauss $\widehat{=}10^{-4}$ T) is also still used.)

Using these two new quantities, Eqns. (5.1) and (5.2) take on the forms

$$\int U\mathrm{d}t = N_\mathrm{J}\,\Delta\Phi\,, \tag{5.5}$$

$$\int U\mathrm{d}t = N_\mathrm{J}A\,\Delta B\,, \tag{5.6}$$

$$U = N_\mathrm{J}\,\dot{\Phi}\,. \tag{5.7}$$

For $N_\mathrm{J} = 1$, i.e. an induction *loop* instead of an induction *coil*, Eq. (5.5) becomes

$$\int U\mathrm{d}t = \Delta\Phi\,,$$

Voltage impulse = change in the magnetic flux; this is formally analogous to the important mechanical equation (Vol. 1, Sect. 5.5):

$$\int F\mathrm{d}t = \Delta p\,,$$

(Mechanical) impulse = change in the momentum.

Making use of Eq. (5.5), we can measure the magnetic flux of a field coil in a very simple way. As in Fig. 5.5, we surround the field coil by an induction coil, switch the current in the field coil on or off, and measure the resulting voltage impulse $\int U\mathrm{d}t$. Then we have $\Phi = \int U\mathrm{d}t/N_\mathrm{J}$, the magnetic flux of the field coil, as sought.

Figure 5.8 shows an example. There, $N_\mathrm{J} = 1$, i.e. instead of an induction *coil* with N_J turns in its winding, a simple induction *loop* is used. From the magnetic flux Φ of the field coil, we obtain its magnetic flux density B (also called the *magnetic induction field*) by dividing Φ by the cross-sectional area of the field coil.

The law of induction makes it possible to readily measure the magnetic field H or the magnetic flux density $B = \mu_0 H$ in *in*homogeneous fields as well: We need only make use of an induction coil or loop of sufficiently small area.

As an example, we measure the magnetic flux density between the flat poles of the electromagnet in Fig. 8.2. We set up a small induction coil J (Fig. 5.9), often called the *probe coil*, perpendicular to the field lines in the region of the field that we want to measure, and connect the ends of the coil to a calibrated ballistic galvanometer. We then observe the voltage impulse when the field is switched on or off, and divide it by the number of 'area turns' $N_\mathrm{J}A$ of the probe coil. We thus find for the electromagnet of Fig. 8.2: $B = 1.5$ V s/m^2 or 1.5 tesla (that is, $H = B/\mu_0 = 1.2\cdot 10^6$ A/m).

C5.5. The introduction of the vector field $\boldsymbol{B}$, which differs from the magnetic field $\boldsymbol{H}$ only by a factor of proportionality μ_0, might seem superfluous here. Indeed, the description of magnetic fields *in vacuum* requires only one of the two field quantities; in general, $\boldsymbol{B}$ is preferred. In some textbooks, the field $\boldsymbol{H}$ is in fact not even mentioned, and $\boldsymbol{B}$ is defined without further ado as *the magnetic field*. However, when matter, in particular magnetic materials are present, the simple relation $\boldsymbol{B} = \mu_0\boldsymbol{H}$ no longer suffices. Then, both of the field quantities are generally required (see Chap. 14).

Video 5.1:
"Induction in Conductors at Rest"
http://tiny.cc/ebggoy
In this video, instead of an induction loop, a coil with N_J turns is used. The data of the setup are: Field coil: $N = 2400$, $l = 0.8\,\mathrm{m}$, diameter $= 5.8\,\mathrm{cm}$, $I = 0.8\,\mathrm{A}$; induction coil: $N_J = 40$; and the ballistic calibration factor of the galvanometer is $B_U = 3.2 \cdot 10^{-5}$ V s/scale division. Note that when the measurement is carried out with the induction coil at the end of the field coil, the resulting voltage impulse is half as large as if the induction coil were inside the homogeneous field. See also Sect. 8.6 and **Exercise 5.1**.

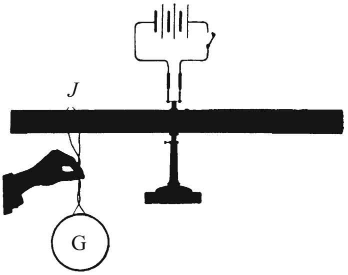


Figure 5.8 Measurement of the magnetic flux Φ of a long coil using an induction loop J ($N_J = 1$) (**Video 5.1**)

Figure 5.9 Probe coil for measuring the magnetic flux density of an electromagnet (one turn with an area of 3 cm^2; this corresponds to $N_J A = 3 \cdot 10^{-4}\,\mathrm{m}^2$ 'area turns')

5.5 Induction in Moving Conductors

In Sects. 5.3 and 5.4, we made use of fixed coils and varied the magnetic field strength only. In the resulting Eq. (5.1), the cross-sectional area A of the field region and the field strength H occurred as equally-important factors. Now, we will describe an experiment belonging to group 3 as defined in Sect. 5.2: We keep the field strength H (or B) constant, but vary the *cross-sectional area* A. Again, we make use of a homogeneous magnetic field. In Fig. 5.10, we are looking parallel to the field lines into a long coil ($H \approx 5000$ A/m). In the circular field of view, we see at the left two metal wires which are bent at right angles. Their ends emerge from the field coil through slits in its left side and are connected to a ballistic galvanometer, calibrated in volt second. At the right, the two horizontal wires are bridged by a *slider* of length L which can slide along them; it can be moved along an arbitrary distance Δx using the handle at right. This changes the area of the loop by $\Delta A = \Delta x L$. At the same time, we observe a voltage impulse. Its magnitude$^{\mathrm{C5.6}}$ is determined experimentally by the galvanometer to be:

$$\int U\,\mathrm{d}t = \mu_0 H \Delta A = B\,\Delta x L\,. \qquad (5.8)$$

C5.6. Here, again, the *sign* is not taken into account. It could however be determined unambiguously from the experiment in Fig. 5.10.

As an application of Eq. (5.8), we measure the horizontal component of the earth's magnetic field in Göttingen. For such measurements, one generally employs a flat induction coil J about as large as a human hand, with several hundred turns in its windings. In this case,

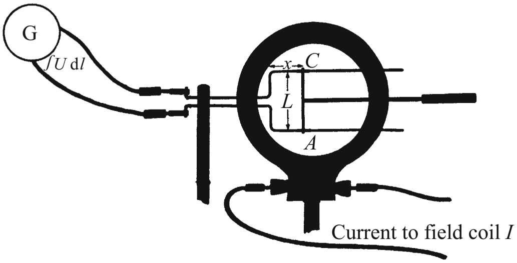


Figure 5.10 Induction in moving conductors; the induction loop contains a *slider* $C-A$ of length L on one side. The magnetic field is perpendicular to the plane of the page, pointing towards the observer. It is produced by a field coil with 10 turns per centimeter of length, i.e. $N/l = 1000/\mathrm{m}$. (The measuring instrument (galvanometer) is outside the magnetic field.) (**Video 5.2**)

Video 5.2:
"Induction in Moving Conductors"
http://tiny.cc/hbggoy
Field coil: $N/l = 1000/\mathrm{m}$, $I = 6\,\mathrm{A}$, $L = 6.5\,\mathrm{cm}$, $\Delta x = 8\,\mathrm{cm}$. As an introduction, the experiment is first shown when turning on the current in the field coil. We then observe the same voltage impulse as in the experiment in **Video 5.1** when the current in the field coil was switched off while the coils remained fixed. (**Exercise 5.2**)

it is not called a probe coil, but rather an *earth inductor*. To produce the voltage impulse, the plane of the coil is oriented vertically (perpendicular to the component to be measured) and the coil is then rotated by an angle of 180°.[C5.7] We obtain $B_{\mathrm{hor}} = 0.2 \cdot 10^{-4}\,\mathrm{Vs/m^2}$ (corresponding to a field strength of $H_{\mathrm{hor}} = 16\,\mathrm{A/m}$).

C5.7. Here, the area of the coil is not changed, but due to the change in its angle, the projection of its area onto the field direction is varied; or, expressed differently: The scalar product $\boldsymbol{B}\cdot\boldsymbol{A}$ changes (in this experiment by $2BA$), and the voltage impulse is proportional to its change (**Exercise 5.3**).

With induction in fixed coils (Sect. 5.3), we were able to maintain a constant voltage U instead of a voltage impulse, at least for a short time. We had only to give the magnetic field a constant rate of change $\dot{H} = \mathrm{d}H/\mathrm{d}t$. We can carry out a similar experiment using induction with moving conductors as described above; we can move the slider along the x direction at a constant velocity $u = \mathrm{d}x/\mathrm{d}t$. Then from Eq. (5.8), we obtain a constant voltage between the ends of the induction loop:

$$U = BuL\,. \tag{5.9}$$

This is a formulation of the law of induction for a *moving conductor* which stays within a region of constant flux density during its motion. Note that in this and the next sections, we use a *fast-responding* galvanometer calibrated in volts, *not* a ballistic galvanometer, to determine the induced voltage U.

> One must avoid bringing the voltmeter or galvanometer, including its lead wires, into the magnetic field and moving it with the slider; it then might not indicate the correct voltage U. More details on this topic will be given in Sect. 7.3.

In Eq. (5.9), the essential quantity is the *velocity* u with which the slider is moving perpendicular to the direction of the field lines and relative to the source of the magnetic field (the field coil). We can demonstrate this in a very dramatic way: We replace the slider wire (Fig. 5.10) by a broad metal ribbon (Fig. 5.11). It consists of an endless band which is moving through the field coil. It permits us to maintain a constant velocity of motion u for an arbitrarily long time and to observe the resulting constant induced voltage.[C5.8]

C5.8. If we use instead a metal disk which can rotate around an axle perpendicular to the plane of the page, we would have constructed a so-called "Barlow's wheel". Peter Barlow (1776–1862), "unipolar inductor", 1823.

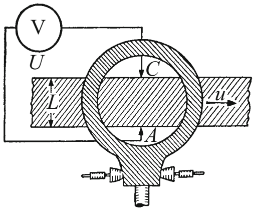


Figure 5.11 Induction in moving conductors; here, as *slider*, we use an endless metal ribbon which is joined to form a loop outside the field coil like the blade of a band saw (the magnetic field is the same as in Fig. 5.10.) (**Exercise 5.4**)

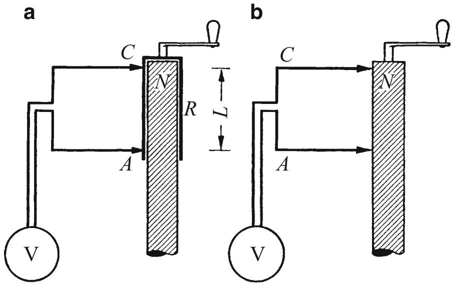


Figure 5.12 Induction of a constant voltage in moving conductors within a radially-symmetric magnetic field (produced by the bar magnets (shaded)). The conductors ("rotors") which are moving relative to the source of the magnetic field consist in part a) of a short piece of metal tubing *R*, and the surface of the magnet itself in part b). They follow circular paths, perpendicular to the field lines.

For qualitative experiments, the simple arrangement sketched in Fig. 5.12 is sufficient. We surround the pole region of a bar magnet loosely by a short piece of metal tubing *R* which can be rotated using a crank. The magnetic field lines pass nearly perpendicular through the walls of this rotor, and its path velocity *u* is perpendicular (like that of the endless band in Fig. 5.11) to the direction of the field lines (Fig. 4.7). The voltage *U* is induced between the two ends of the tube (length *L*). The usual name for this arrangement is *unipolar induction*.

We could also connect the tube rigidly to the bar magnet and let it rotate around its long axis. Then the tube would become superfluous; it would be sufficient to let the sliding contacts slip over the surface of the bar itself – see Fig. 5.12b. Keep in mind that the field lines are not fixed on the surface of the bar magnet like the bristles of a brush![C5.9]

C5.9. For a detailed description of these experiments, see J. W. Then, *American Journal of Physics* **30**, 411 (1962).

5.6 The Most General Form of the Law of Induction

A general form of the law of induction can be found if we relinquish our requirement of a *homogeneous and time-independent* magnetic field and also take the sign found in the experiments into account[2]. Then one can define experimentally the voltage induced in a loop:C5.10

C5.10. Equation (5.10) holds for all induction experiments in the three groups described in Sect. 5.2.

$$U = -\frac{\mathrm{d}}{\mathrm{d}t}\int \boldsymbol{B}\cdot \mathrm{d}\boldsymbol{A}\,. \tag{5.10}$$

The integral on the right-hand side can be decomposed into two parts: A time-dependent change of the magnetic flux density $\boldsymbol{B}$, and a spatial change of the curve around the perimeter s of the integrated area A (the induction loop) at whose ends the voltage is measured. We find:C5.11

C5.11. For the derivation of Eq. (5.11), see for example P. Lorrain and D. R. Corson, "*Electromagnetic Fields and Waves*", 2nd edition, W. H. Freeman & Co., San Francisco 1970, Chap. 8.

$$U = \oint_{s(A)} \boldsymbol{E}\cdot \mathrm{d}\boldsymbol{s} = -\int_{A} \frac{\mathrm{d}\boldsymbol{B}}{\mathrm{d}t}\cdot \mathrm{d}\boldsymbol{A} + \oint_{s(A)} (\boldsymbol{u}\times\boldsymbol{B})\cdot \mathrm{d}\boldsymbol{s}\,. \tag{5.11}$$

(Here, $\boldsymbol{u}$ is the velocity with which a line element $\mathrm{d}\boldsymbol{s}$ of the perimeter curve is moving relative to the frame of reference in which U and $\boldsymbol{B}$ are *measured*).

The direction of the surface element in the first term and the direction of the integration along the path integrals when combined have to yield a right-hand screw. With this formulation, the law of induction has the correct sign (see Sect. 6.1).C5.12

C5.12. The velocity $\boldsymbol{u}$ is by no means necessarily constant along the perimeter curve (examples: experiments of group 3, or also Fig. 5.10).

The attribution of the induced voltage to the two parts of the integral can change when the frame of reference is changed. Some examples of this: In experiments from group 2 in Sect. 5.2, only the first term on the right is present when the position of the induction coil is used as the frame of reference. However, if the position of the field coil defines the frame of reference, both terms have to be taken into account (of course, then $\mathrm{d}\boldsymbol{B}/\mathrm{d}t$ has a different value). Or if in Fig. 5.6, we let the induction coil rotate and use the field coil as the frame of reference, then only the second term on the right of Eq. (5.11) contributes. But within the frame of reference of the induction coil, only the first term contributes (Exercise 5.6). In contrast, in Fig. 5.11, only the second term is present, since there, $\mathrm{d}\boldsymbol{B}/\mathrm{d}t = 0$ (Eq. 5.9).C5.13

C5.13. The description of this experiment in the frame of reference of the metal band is most readily formulated using the theory of relativity (see Sect. 6.1).

Equation (5.10), while often used, is sometimes not very suitable for practical calculations, e.g. when the area over which the integration is carried out is not uniquely determined, as in Fig. 5.11. In such cases, one should return to Eq. (5.1) or to Eq. (5.8).

Exercises

5.1 Using the information given in Video 5.1, calculate the expected voltage impulse $\int U\,\mathrm{d}t$ and compare it with what was observed in the video. (Sect. 5.4)

[2] Compare the small-print paragraph near the end of Sect. 5.3 for remarks about the sign.

5.2 With the information given in Video 5.2, compute a) the velocity u with which the slider in Fig. 5.10 must be moved in order to obtain a voltage of $U = 1\,\text{mV}$ from the induction loop (measured with a voltmeter having a rapid response time); and b) the expected voltage impulse $\int U\,\text{d}t$ which would be observed with a slowly-responding galvanometer, cf. Sect. 5.2. Compare the latter with what was observed in the video. (Sect. 5.5)

5.3 The measurement of the earth's magnetic field in the Göttingen lecture hall: An earth inductor with a cross-sectional area of $10^3\,\text{cm}^2$ and 200 turns in its windings is oriented with the aid of a compass needle so that its axis points in the magnetic north-south direction. Then the axis is moved to a horizontal orientation. If the earth inductor is now rotated around a diameter by 180°, we observe a voltage impulse of $10^{-3}\,\text{V s}$. a) Determine from this result the horizontal component B_h of the earth's magnetic field. b) In a second experiment, the axis of the earth inductor is oriented vertically. When it is rotated around a diameter by 180°, now a voltage impulse of $2.25 \cdot 10^{-2}\,\text{V s}$ is observed. Find the angle φ between the direction of the earth's field and the horizontal (called the angle of inclination). (Sect. 5.5)

5.4 The endless metal band in Fig. 5.11 is 10 cm wide and takes the form of a circle with a diameter of 1 m. The magnetic field is the homogeneous field at the center of a 50 cm long field coil with 1000 turns, carrying a current of 1.2 A. Find the frequency ν at which the band would have to rotate in order to generate a voltage of 1 mV between the slip contacts A and C. (Sect. 5.5)

5.5 An aircraft with a wingspan of 50 m is flying at a constant altitude with a velocity of 960 km/h. The pilot has a voltmeter which is attached to the wingtips with insulated wire leads. The wings and fuselage of the aircraft are electrically conducting and are connected together. The vertical component of the earth's magnetic field is $6 \cdot 10^{-5}\,\text{T}$. What voltage will be indicated by the voltmeter? (Sects. 5.5 and 7.3)

5.6 Making use of Eq. (5.10), compute the voltage U which will be induced in the rotating coil from Sect. 5.3 (Fig. 5.6), if it is rotated around an axis perpendicular to the plane of the page. The cross-sectional area of the coil is A, it has N turns in its windings, and its rotational frequency is ν. (Sect. 5.6)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_5) contains supplementary material, which is available to authorized users.

The Relation Between Electric and Magnetic Fields

6

6.1 Detailed Treatment of Induction. MAXWELL's Second Law

We return to the first experiment (in group 1) of Sect. 5.2, and reconsider induction in the simplest possible case: An induction coil with only *one* turn, i.e. an induction loop, is supposed to enclose a *time-dependent* magnetic field of flux density B along some arbitrary curve s with a cross-sectional area A (Fig. 6.1). Then at the ends of the loop, we observe the induced voltage (without taking its sign into account):

$$U = \dot{B}A\,. \tag{5.2}$$

This experimental result can be interpreted in a more profound sense as follows: *The conductor, the single loop of wire, is quite unimportant and insignificant. The actual phenomenon does not depend on the accidental presence of the wire loop. It consists of the appearance of electric field lines along closed paths around the time-dependent magnetic field* (Fig. 6.2).[C6.1] Electric field lines in the form of closed circles are something quite new and completely unexpected. Up to now, we have encountered only electric field lines with ends; at their ends were electric charges.

Let us continue our investigation: *The wire loop is merely an indicator to detect the electric field.* Along its length, it measures the

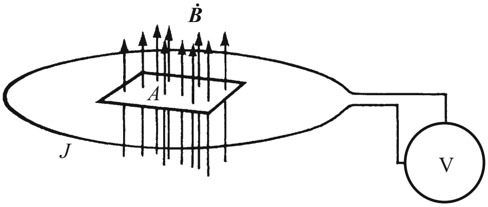


Figure 6.1 Schematic of an induction experiment with an induction coil of only one turn ($N_\mathrm{J} = 1$). The vector field $\dot{\boldsymbol{B}}$ is the time derivative of the vector field $\boldsymbol{B}$. The instrument represents the induced voltage.

C6.1. A very impressive demonstration of such field lines on closed paths is given by the *betatron*; this is a device in which electrons can be accelerated, and it is described in many introductory physics textbooks. However, to understand it properly, one needs to know about the LORENTZ force, which we introduce in the following chapter. (See e.g. *The Feynman Lectures on Physics*, Addison-Wesley 1964, Vol. II, Chap. 17 (online at http://www.feynmanlectures.caltech.edu/); cf. also Fig. 11.11).

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_6

Figure 6.2 The deeper significance of the phenomenon of induction

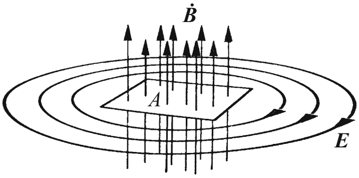
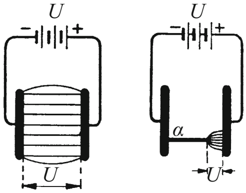


Figure 6.3 The function of the wire loop in the induction experiment



path integral of the electric field strength $\boldsymbol{E}$, i.e. the induced voltage $U = \int \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{s}$. It has no other function than the wire α in the schematic in Fig. 6.3: The wire is a conductor and causes the electric field to decay in its interior. The charges move to its ends and thus the voltage acting all along its length is compressed into the remaining gap.

In the electric fields which we have thus far investigated, the path integral of the field $\boldsymbol{E}$ along a closed loop was always zero. It was, independently of the exact path s followed, just equal to the voltage between the beginning and the end of the path (see Eq. (2.3)). It was therefore zero whenever the beginning and the end approached each other closely; at the limit of a vanishing gap, it vanished exactly.

The situation is different with the fields that we encounter in this induction experiment, which form endless loops; here, the electric voltage along a closed path has a finite value. Furthermore, when the number of turns is increased to N_J, so that the magnetic field is circled N_J times, the voltage increases correspondingly (Eq. (5.2)).

In this way of looking at the experiment, the electric field that is induced during an induction process is the primary phenomenon. The observed voltage is the path integral of that electric field $\boldsymbol{E}$. It has the value (compare Eq. (2.3)):

$$U = \oint \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{s}\,. \tag{6.1}$$

Then, taking the sign into account, Eq. (5.2) takes the following form:[C6.2]

$$\oint \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{s} = -\dot{B}A = -\mu_0\dot{H}A\,. \tag{6.2}$$

This equation yields the *electric* field that is produced by the change of a *magnetic* field. It summarizes the essential content of the second of the four MAXWELL*'s laws*.

C6.2. Written completely and consistently in vector form, Eq. (6.2) becomes

$$\oint_{s(A)} \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{s} = -\frac{\mathrm{d}}{\mathrm{d}t}\int_A \boldsymbol{B}\cdot\mathrm{d}\boldsymbol{A}\,.$$

As already agreed upon in Sect. 5.6, the path of integration s encloses the area A in such a way that $\mathrm{d}\boldsymbol{s}$ is positive (it curves in a “right-hand” sense when one looks along the direction of the surface-element vector $\boldsymbol{A}$). The *sign* will be treated in detail once more in Sect. 8.3; it is given by LENZ*'s law*.

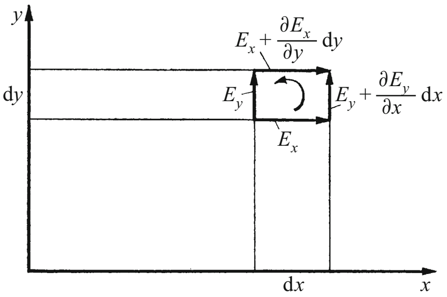

y
$E_x + \frac{\partial E_x}{\partial y}\,\mathrm{d}y$
$\mathrm{d}y$
E_y
$E_y + \frac{\partial E_y}{\partial x}\,\mathrm{d}x$
E_x
$\mathrm{d}x$
x

Figure 6.4 Taking the path integral of the electric field $\boldsymbol{E}$ along the perimeter of a surface element $\mathrm{d}x\,\mathrm{d}y$; the z axis is perpendicular to the plane of the page and points towards the reader (right-handed coordinate system). The path of integration (curved arrow) is in the clockwise sense (to the right) as seen along the z direction (similar to Fig. 6.15).

The equation can be written in a differential form and can then be applied to arbitrary inhomogeneous magnetic fields. Its derivation from Eq. (6.2) can be accomplished by taking the path integral along the perimeter of a surface element $\mathrm{d}x\,\mathrm{d}y$. This computation is explained in Fig. 6.4. One thus obtains (taking the sign of $\dot{\boldsymbol{B}}$ in Fig. 6.2 into account!):

$$\frac{\partial E_\mathrm{y}}{\partial x} - \frac{\partial E_\mathrm{x}}{\partial y} = -\dot{B}_z$$

or, after taking account of the other components, in vector notation,

$$\operatorname{curl}\boldsymbol{E} = -\dot{\boldsymbol{B}} = -\mu_0\,\dot{\boldsymbol{H}}\,. \tag{6.3}$$

In words: At each point within a magnetic field, a time-dependent change in the field strength or direction produces an electric field. The latter is a *vortex field*, and the curl of the field $\boldsymbol{E}$ is equal to the negative rate of change of the field $\boldsymbol{B}$. (For the definition of the "*curl*", see Vol. 1, Sect. 10.7).

The third of MAXWELL's laws contains another, analogous relation between the two fields, but with the roles of $\boldsymbol{E}$ and $\boldsymbol{H}$ reversed. Its experimental derivation is our next goal.

6.2 Measuring Magnetic Potential Differences

We know from ROWLAND's experiment (Sect. 4.3) that every motion of electric charges represents an electric current, and every current produces a magnetic field. We can indeed measure the magnetic field using the current. But we still lack a general experimental framework for the relation between the current and the resulting magnetic field, as described in MAXWELL*'s third law*. We can arrive at it by measuring the *magnetic potential difference*.

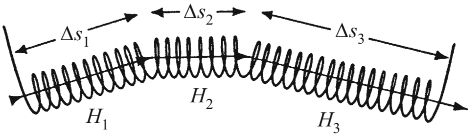


Figure 6.5 Schematic of a magnetic potentiometer (A. P. CHATTOCK, *Phil. Mag.* **24**, 94 (1887); W. ROGOWSKI and W. STEINHAUS, *Arch. f. Elektrotech.* **1**, 141 (1912))

In an *electric* field, we found that the electric potential difference or voltage is given by the path integral of the electric field:

$$U = \int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} \,. \tag{2.3}$$

Its unit, as usual, is 1 volt.

In a corresponding fashion, in a *magnetic* field, we can define the path integral of the field $\boldsymbol{H}$ as the *magnetic potential difference*:

$$U_{\text{mag}} = \int \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{s} \,. \tag{6.4}$$

Its unit can be seen to be 1 ampere.

The magnetic potential difference can be measured with a very simple instrument, the *magnetic potentiometer*. *A magnetic potentiometer is in principle simply a very long induction coil*, wound for example on a long strap. It is wound with two layers and the leads are brought out at the center of the outer layer of windings (Fig. 6.6). (A coil with only a single layer would, in addition to the long coil as intended, also represent a large, flat induction coil formed by a single loop of helically-coiled wire).

We want to elucidate the operation of this magnetic potentiometer: The magnetic potential difference is to be measured along a path s. This path is divided up in Fig. 6.5 into a broken curve with segments (path elements) Δs_1, Δs_2, $\ldots$, Δs_n. Let us denote the components of the field in the direction of the path elements Δs as H_1, H_2, $\ldots$, H_n. The magnetic potentiometer extends along the whole path s. It has N turns in its winding and the length l. Its n-th path element has the length Δs_n. Then there are $N_\mathrm{n} = N \cdot \Delta s_\mathrm{n}/l$ turns in this element. If the field H is increasing or decreasing, then a certain voltage impulse will be induced in the magnetic potentiometer, $\int U\,\mathrm{d}t$ (Eq. (5.1)). This is composed of the sum of the contributions from each of the path elements; thus, if A is the (rectangular) cross-sectional area of the windings of the potentiometer (the sign holds for a decreasing field H), the voltage impulse will be:

$$\int U\,\mathrm{d}t = \mu_0 A H_1 N \Delta s_1/l + \mu_0 A H_2 N \Delta s_2/l \cdots + \mu_0 A H_\mathrm{n} N \Delta s_\mathrm{n}/l\,, \tag{6.5}$$

$$\int U\,\mathrm{d}t = \mu_0 A N (H_1 \Delta s_1 + H_2 \Delta s_2 + H_3 \Delta s_3 \cdots + H_\mathrm{n} \Delta s_\mathrm{n})/l\,, \tag{6.6}$$

$$\int U\,\mathrm{d}t = \mu_0 A N \left(\int \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{s} \right) /l = \mu_0 A N\, U_\mathrm{mag}/l\,, \tag{6.7}$$

$$U_\mathrm{mag} = \frac{l}{\mu_0 N A} \int U\,\mathrm{d}t\,. \tag{6.8}$$

This induced voltage impulse, measured for example in volt second, after multiplication by the the apparatus constant $l/(\mu_0 NA)$, yields directly the magnetic potential difference as sought, in ampere. The apparatus constant need be determined only once; A, l and N by direct measurement, while $\mu_0 = 1.257 \cdot 10^{-6}$ V s/A m.

We employ a magnetic potentiometer which is 1.2 m long. Its apparatus constant has the value $5 \cdot 10^5$ A/V s. The coil has 9600 turns, each with a cross-sectional area of 2 cm^2. The induced voltage impulse is measured using the ballistic galvanometer described in Sect. 5.2; it can be calibrated as shown in Fig. 5.3 (J is the magnetic potentiometer). (cf. **Video 6.1**)

Video 6.1:
"The Magnetic potentiometer"
http://tiny.cc/obggoy
The potentiometer shown in the video has practically the same dimensions. Its apparatus constant is $4.58 \cdot 10^5$ A/V s. The ballistic calibration factor of the galvanometer has the value $B_\mathrm{U} = 1.1 \cdot 10^{-4}$ V s/scale division (**Exercise 6.1**).

6.3 The Magnetic Potential Difference of a Current. Applications

The operation of the magnetic potentiometer is explained in Fig. 6.6. The magnetic potential difference U_mag of a field coil is to be measured between the points 1 and 2 along the path 1 → 2. The potentiometer is shaped so that it corresponds to the path from 1 to 2; then the magnetic field is varied by switching the current in the field coil between zero and its maximum value, and the resulting induced voltage impulse is observed with the galvanometer.[C6.3] In this way, we reach the following conclusions:

1. Along an open path (Fig. 6.6), the magnetic potential difference depends only on the position of the end points 1 and 2, and not on the shape of the path itself. The path can even include loops, as long as they do not enclose any currents.

2. In Fig. 6.7, the path which the potentiometer follows is closed, and it encloses no currents. The resulting magnetic potential difference is zero.

3. In Fig. 6.8, the path defined by the potentiometer encloses a current I once within a closed loop. The magnetic potential difference

C6.3. We see here that the operation of the magnetic potentiometer involves simply an application of the law of induction (Eq. (5.5)). The magnetic potentiometer is just an induction coil with a particular shape with which voltage impulses $\int U\,\mathrm{d}t$ can be measured.

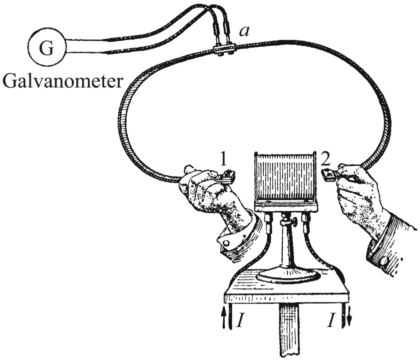


Figure 6.6 Operation of the magnetic potentiometer

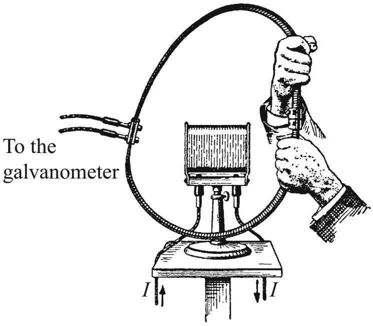


Figure 6.7 A closed path for the magnetic potentiometer which encloses no currents

U_{mag} is again independent of the exact shape of the path (circular, rectangular etc.).

4. Quantitatively, we find in Fig. 6.8 the magnetic potential difference equal to the current I in the conductor which is enclosed in the path

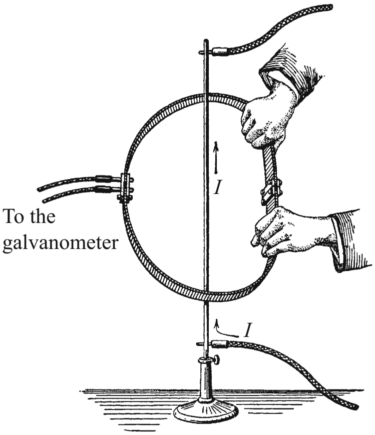


Figure 6.8 Enclosing a current one time with a magnetic potentiometer ($I = 50$ to 100 A, a 2-volt storage battery is sufficient to provide the current) **(Video 6.1)**

Video 6.1:
"The Magnetic Potentiometer"
http://tiny.cc/obggoy
In the video, the potentiometer which encloses a current is moved relative to the current-carrying conductor; but no voltage impulse is observed as a result of this motion.

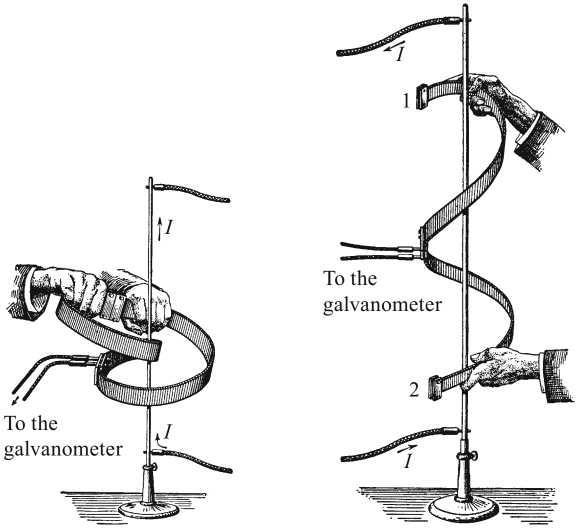


Figure 6.9 A current enclosed twice within the path of a magnetic potentiometer – at left, within a closed loop, and at right within an open curve whose upper and lower ends are one above the other vertically **(Video 6.1)**

of the potentiometer. We have

$$U_{\text{mag}} = \oint \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{s} = I\,. \tag{6.9}$$

This equation is known as "AMPERE's law".

> A numerical example: $I = 83\,\text{A}$. A deflection of the ballistic galvanometer of 12 cm corresponds to $\int U\mathrm{d}t = 1.7 \cdot 10^{-4}\,\text{V s}$. Multiplication by the apparatus constant of the magnetic potentiometer, i.e. by $5 \cdot 10^{5}$ A/V s, yields the magnetic potential difference $U_{\text{mag}} = 1.7 \cdot 10^{-4}\,\text{V s} \cdot 5 \cdot 10^{5}\,\text{A/V s} = 85\,\text{A}$.

5. In Fig. 6.9 (left), the path of the potentiometer encloses the current twice (double loop). The magnetic potential difference is then doubled. Continuing in this manner, we find for an N-fold enclosure of the current I that the measured magnetic potential difference becomes

$$U_{\text{mag}} = \int \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{s} = NI\,. \tag{6.10}$$

6. In Fig. 6.9 (left), the current I is enclosed twice within the path of the potentiometer, and its beginning and end points are held together. This is however not necessary; the potentiometer could just as well be in the form of an open helix with N turns enclosing the current N times and with open ends (Fig. 6.9 (right)).

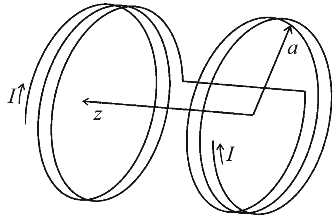


Figure 6.10 HELMHOLTZ coils

Summary: *The magnetic potential difference along an arbitrary curve is simply equal to the strength of the current if the latter is enclosed one time within the curve. When it is enclosed N times, the potential difference becomes N times the current.* This result is expressed by Eq. (6.10) in compact form.

To remind us of these important facts, the following three examples of applications can be helpful:

1. *The homogeneous magnetic field of a long coil* (Fig. 6.11). The magnetic potentiometer is threaded through the coil and closed outside it along an arbitrary path. Its path thus encloses once each N wires carrying the current I. Therefore, the magnetic potential difference along the whole path is $U_{\text{mag}} = N\,I$. U_{mag} is composed of two additive terms, $U_{\text{mag,i}}$ and $U_{\text{mag,o}}$ ('inside' and 'outside'). Within the coil, the magnetic field is homogeneous, apart from the short pole regions (coil ends), and its field strength H is constant.C6.4

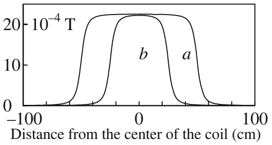
Thus, $U_{\text{mag,i}} = Hl$. We find that the term due to the region outside the coil, $U_{\text{mag,o}}$, is negligibly small compared to $U_{\text{mag,i}}$ (Fig. 6.11b). Then we have

$$Hl = NI \quad \text{or} \quad H = \frac{NI}{l}\,.$$

This is just the same as Eq. (4.1) from Sect. 4.2. It proves here to have been a special case of the general equation (6.10)[1]

2. *The magnetic field $H(r)$ at a distance r from a current-carrying straight wire*. The magnetic potential difference along one of its circular field

C6.4. In order to get an impression of the homogeneity of the field of a long coil, the following figure shows the strength of the axial field, i.e. the "field profile" of long coils.



The longitudinal component of the $\boldsymbol{B}$ field was computed along the axis of two coils, both with $N\,I/l = 1800$ A/m. The field strength in their interior at the center is $B = 22.5 \cdot 10^{-4}$ T. Curve a): length/diameter = 100 cm/10 cm; b): 50 cm/10 cm (the same geometry as in Fig. 4.4). The decrease of B in the neighborhood of the ends of the coil is practically independent of the length of the coil. (The calculation was performed by J. A. Crittenden, Cornell University.)

[1] We mention without derivation that the field strength at the center of a cylindrical coil (solenoid) of radius r and length l is given by

$$H = \frac{NI}{l}\,\frac{l}{\sqrt{4r^2 + l^2}}\,. \tag{6.11}$$

For $r \ll l$, we see that Eq. (4.1) applies, and at the center of a current-carrying circular ring ($N = 1$, $l = 0$), we find

$$H = I/2r\,. \tag{6.12}$$

An often-used arrangement for producing homogeneous fields consists of two such circular rings, called HELMHOLTZ *coils* (Fig. 6.10).

If the spacing of the two coils (each with radius a and N turns in their windings) is equal to their radius, then the field H along the z axis ($z = 0$ at the center between the coils) is given by

$$H = H_0(1 - 1.15(z/a)^4 + \ldots) \quad \text{with} \quad H_0 = 0.716NI/a\,,$$

and thus, along the z axis for $|z| \le 0.1\,|a|$, it is constant to within 10^{-4}.

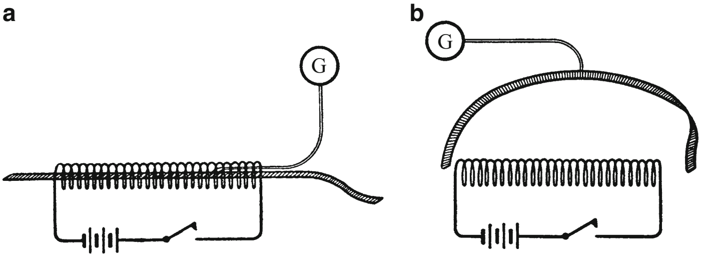


Figure 6.11 The distribution of the magnetic potential difference in the field of a long coil. The coil has 900 turns in its windings, a length of 0.5 m and a diameter of 0.1 m. A current of 1 A passed through the coil produces a magnetic field $H = 1800$ A/m. a): The magnetic potentiometer is passed through the entire length of the field coil. Switching the current off or on yields a voltage impulse of $1.7 \cdot 10^{-3}$ V s, i.e. from Eq. (6.8), $U_{\text{mag}} = 850$ A. The length and position of the ends of the potentiometer coil outside the field coil are practically unimportant. Thus, the field outside the field coil makes no significant contribution to the magnetic potential difference.
b): The potentiometer follows an arbitrary path completely outside the field coil. The voltage impulse which is induced in it is only about $9 \cdot 10^{-5}$ V s. U_{mag} thus has a value in the region outside the field coil of only around 45 A, and can therefore be neglected relative to the magnetic potential difference of 850 A measured within the field coil. The path integral $\int \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{s}$ for the outside region is indeed practically negligible, even for this field coil which is not really very long. **(Video 6.1)** (**Exercise 6.2**)

Video 6.1:
"The Magnetic Potentiometer"
http://tiny.cc/obggoy
In this experiment, the magnetic potential difference in the region outside the coil is ≈ 10 % of that measured within the coil. The data of the field coil are: Windings: $N = 4300$, length $l = 40$ cm, diameter = 11 cm, current $I = 0.15$ A.
$U_{\text{B}} = 3.2 \cdot 10^{-5}$ V s/scale division.

lines (Fig. 1.4) with a radius r can be found from Eq. (6.10), taking the symmetry of the problem into account:[C6.5]

$$U_{\text{mag}} = 2\pi r H(r) = I\,,$$

and thus

$$H(r) = \frac{1}{2\pi}\,\frac{I}{r}\,. \tag{6.13}$$

The field is directed tangentially, and its strength is determined only by r (for its *sign*, see Fig. 4.10).

3. *Magnetic potential difference measurements in the magnetic fields of permanent magnets*. Our treatment has emphasized the essential similarity of the magnetic fields of current-carrying conductors and those of permanent magnets. This can be reinforced once more by measurements with the magnetic potentiometer. In Fig. 6.12, the magnetic potential difference between the poles of a horseshoe magnet is being determined. For these measurements, the magnet is pulled away from the potentiometer rapidly. The potential difference is once more found to be completely independent of the path which is investigated with the potentiometer (i.e. the shape of the curve which the potentiometer represents). On a closed loop, it is always zero. The potentiometer can of course not enclose individual molecular currents; it would have to be threaded through single molecules for that! Every hole that might be bored through a permanent magnet however passes not through the individual molecules, but rather between them.

C6.5. **"taking the symmetry into account"** means here that one can find no reason why the tangential component of the magnetic field on a concentric circle (of radius r) around the current path (wire) should not be constant. It must therefore be constant. The fact that the radial and axial components are both zero can however not be concluded from symmetry alone; it is instead an experimental finding (the absence of a radial component can be clearly seen in Fig. 1.4, for example).

In a few cases with simple geometry in which the strength of the magnetic field along the magnetic potentiometer is constant, we could use

Video 6.1:
"Magnetic Potentiometer"
http://tiny.cc/obggoy

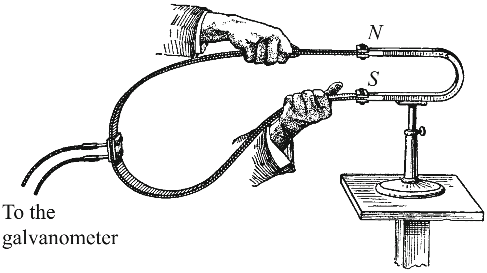


Figure 6.12
A magnetic potentiometer in the field of a permanent magnet. The experimental setup corresponds to the one in Fig. 6.11b. **(Video 6.1)**

the potentiometer to determine the field strength, for example in the case of a single current-carrying wire. The importance of this experiment is however much more far-reaching. The measurement setup as shown in Fig. 6.8 corresponds to the one in Fig. 5.5: The current density[2] $\boldsymbol{j}$ corresponds to the field $\dot{\boldsymbol{B}}$ and the magnetic potential difference $\oint \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{s}$ corresponds to the voltage $\oint \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s}$. Experiments have shown that the magnetic potential difference is equal to the enclosed field of the current density $\boldsymbol{j}$. It then follows, in analogy to the derivation given in Sect. 6.1, that the relation between $\boldsymbol{H}$ and $\boldsymbol{j}$ can be given in the form of a differential expression as:

$$\operatorname{curl} \boldsymbol{H} = \boldsymbol{j}\,, \tag{6.14}$$

where the sign on the right side is to be determined by comparison with Fig. 4.10. This equation is AMPERE's law in differential form. It is a part of the third of MAXWELL's laws, which will be treated in the following section.

6.4 The Displacement Current and MAXWELL's Third Law

The experimental result that $\oint \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{s} = I$ (Eq. 6.9) was generalized by MAXWELL in an audacious manner. His train of thought can be explained by referring to Fig. 6.13. A condenser is discharged through an external circuit. During its discharging, a current I flows through its leads, and within the condenser, its electric field $\boldsymbol{E}$ is decaying at the same time. The current I in the wires is surrounded by concentric circular magnetic field lines. We imagine that this figure could contain field lines around all the wires in the circuit. Then we can say, roughly but unmistakably, that the whole wire is surrounded by a "tube" of magnetic field lines. This tube ends at both sides of the condenser, where the wires meet the plates (electrodes) of the condenser. MAXWELL maintained, on the contrary, that the tube of magnetic field lines has no ends; it forms a closed torus:

[2] The current density j is the current per cross-sectional area, $j = \mathrm{d}I/\mathrm{d}A$; or, more generally, in vector notation, $I = \int \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{A}$.

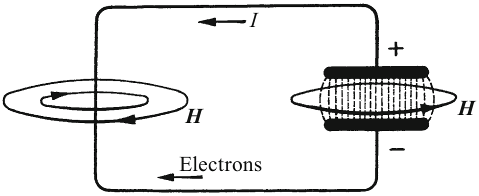
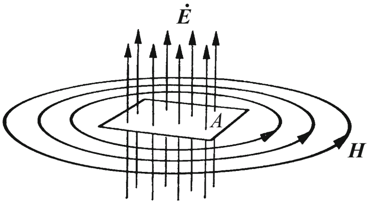


Figure 6.13 Schematic drawing of the magnetic field from a conduction current and a displacement current (I corresponds to the conventional current direction from + to −)

Figure 6.14 Schematic drawing of the magnetic field of a *displacement current*. The vector field $\dot{\boldsymbol{E}}$ is the time derivative of the vector field $\boldsymbol{E}$ (the arrows show the direction of a displacement current I_v which is pointing upwards).



The varying electric field within the condenser is also *surrounded by circular magnetic field lines*. Such circular magnetic field lines are however one of the principal characteristics of an *electric current*. The corresponding current is referred to – somewhat strangely – as the *displacement current*. All of the usual meanings of the word current, 'flowing' or 'streaming' in analogy to a current of water no longer apply to this 'current'. *The word* displacement current *here indeed refers only to the fact that an electric field is changing with time* (Fig. 6.14).

Following the introduction of this new type of current, we can say that: *In nature, there are only closed current loops*. In a conductor, these are conduction currents, and in an electric field (e.g. of a condenser), they are displacement currents. Electric currents have no spatial ends or beginnings. At the end of the conduction current, the displacement current begins, and *vice versa*.

Like every current, the displacement current can be measured in *ampere*. On the other hand, it is supposed to be a quantity which is determined by the time derivative of an electric field. The latter field would therefore have to have the unit *ampere second*. This is the case for the product

$$\text{Displacement density}\,D \\ \cdot\,\text{Cross-sectional area}\,A\text{ of the field} = DA = \varepsilon_0 EA$$

(for example: D in $\mathrm{A\,s/m^2}$, A in $\mathrm{m^2}$, E in $\mathrm{V/m}$, DA in $\mathrm{A\,s}$, $\varepsilon_0 = 8.854 \cdot 10^{-12}\,\mathrm{A\,s/V\,m}$).

The displacement density D (Sect. 2.13) is related to the electric field E via $D = \varepsilon_0 E$ (Eq. 2.5). We denote the rate of change of D and E

again with a dot above the field symbol, i.e. $\dot{D} = \mathrm{d}D/\mathrm{d}t$ and $\dot{E} = \mathrm{d}E/\mathrm{d}t$. Then we obtain the displacement current

$$I_\mathrm{v} = \dot{D}A = \varepsilon_0 \dot{E} A\,. \tag{6.15}$$

The quantity $\dot{D} = I_\mathrm{v}/A$ is also called the "displacement current density".

So much for the measurement of the displacement current. AMPÈRE's law

$$\oint \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{s} = I \tag{6.9}$$

was originally discovered through experiments with the conduction current. MAXWELL extended this law by a term containing the *displacement current*, written as[C6.6]

$$\oint \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{s} = \dot{D}A = \varepsilon_0 \dot{E} A\,. \tag{6.16}$$

C6.6. Written completely in vector notation, Eq. (6.16) becomes

$$\oint\limits_{s(A)} \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{s} = \varepsilon_0 \frac{\mathrm{d}}{\mathrm{d}t} \int\limits_{A} \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{A}\,.$$

Here, the path of integration s encloses the area A in such a way that d$\boldsymbol{s}$ is positive (i.e. it curves clockwise) if one is looking in the direction of the surface-normal vector $\boldsymbol{A}$.

This equation yields the *magnetic* field which is produced by the change in an *electric* field.

The relation described by Eq. (6.16) can be obtained as a differential equation by referring to Fig. 6.15; its derivation is similar to that of Eq. (6.3), above. We have to compute the path integral of $\boldsymbol{H}$ along the perimeter of a surface element dx dy; we thus obtain

$$\frac{\partial H_\mathrm{y}}{\partial x} - \frac{\partial H_\mathrm{x}}{\partial y} = \dot{D}_\mathrm{z} = \varepsilon_0 \dot{E}_\mathrm{z}\,, \tag{6.17}$$

or, after including the other components in vector notation,

$$\operatorname{curl} \boldsymbol{H} = \dot{\boldsymbol{D}} = \varepsilon_0 \dot{\boldsymbol{E}}\,. \tag{6.18}$$

In words: At every point within an electric field, a change of the field with time produces a magnetic field. The latter is a vortex field, and the curl of this magnetic field is equal to the time derivative of the displacement current density. We have assumed here that the surface element dx dy is penetrated by only a displacement current. If in addition a conduction current I passes through the surface element, then on the right side of the equation, its current density $j = \frac{\mathrm{d}I}{\mathrm{d}x\,\mathrm{d}y}$ must be added as a second term.

We have thus obtained MAXWELL's complete third law:

$$\operatorname{curl} \boldsymbol{H} = \boldsymbol{j} + \dot{\boldsymbol{D}} = \boldsymbol{j} + \varepsilon_0 \dot{\boldsymbol{E}}\,. \tag{6.19}$$

Unfortunately, we cannot simply make the magnetic field lines from the displacement current visible in Fig. 6.13 like those of a conduction current by using iron filings. That would be a didactically very expedient, direct demonstration of the validity of the second term on the right side of Eq. (6.19). However, for technical reasons, in electric fields with long field lines, we cannot produce a displacement

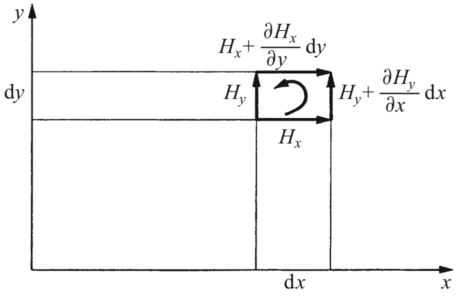

y
$H_x + \frac{\partial H_x}{\partial y}\,\mathrm{d}y$
dy
H_y
$H_y + \frac{\partial H_y}{\partial x}\,\mathrm{d}x$
H_x
dx
x

Figure 6.15 Taking the path integral of the magnetic field $\boldsymbol{H}$ along the perimeter of a surface element dx dy; the z axis points out of the plane of the page (right-hand coordinate system) and the sense of the integration (curved arrow) is clockwise as seen in the z direction (as in Fig. 6.4)

current of sufficient strength. But such a demonstration would in any case not prove that the origin of the magnetic field was to be found in the displacement current; one could always claim that the magnetic field observed in Fig. 6.13 was due to the conduction current in the leads to the condenser plates.

A real proof of the origin of the magnetic field from the displacement current can be obtained only by referring to circular, closed electric field lines. We will do this in Chap. 12, by the detection of freely propagating *electromagnetic waves*. Until then, the magnetic field of the displacement current remains only a plausible hypothesis.

6.5 MAXWELL's Fourth Law

At this point, we want to introduce the last of the four MAXWELL equations. In integral form, it is given by:

$$\oint \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{A} = 0\,, \tag{6.20}$$

i.e. the magnetic flux (Eq. (5.3)) through a closed surface, integrated over the whole surface area, *vanishes*. In differential form, this becomes

$$\operatorname{div} \boldsymbol{B} = 0\,.$$

These equations are analogues of Eqns. (2.9) and (2.8) in Chap. 2, which related electric fields to electric charges. The observation that magnetic field lines have no beginnings and no ends, as shown by the experiments in Chap. 4, means that there is no magnetic analogue of electric charges, i.e. *there are no magnetic charges*. This explains the zero on the right-hand side of the two equations above.

We summarize MAXWELL's four laws (also known as the *Maxwell equations*); they describe electric and magnetic fields in vacuum (and thus practically also in air). Note that in vacuum, $\boldsymbol{D} = \varepsilon_0 \boldsymbol{E}$ and $\boldsymbol{B} = \mu_0 \boldsymbol{H}$. The four laws in differential form (for $\boldsymbol{E}$ and $\boldsymbol{B}$) are:

$$\operatorname{div} \boldsymbol{E} = \frac{\varrho}{\varepsilon_0}\,, \tag{6.21}$$

$$\operatorname{curl} \boldsymbol{E} = -\dot{\boldsymbol{B}}\,, \tag{6.22}$$

$$\operatorname{curl} \boldsymbol{B} = \mu_0 \boldsymbol{j} + \mu_0 \varepsilon_0 \dot{\boldsymbol{E}}\,, \tag{6.23}$$

$$\operatorname{div} \boldsymbol{B} = 0\,. \tag{6.24}$$

The experimental justification of these equations (with the exception of the displacement current density in the third equation) has been given in the preceding sections (for example, in Sect. 2.14 for Eq. (6.21), Sect. 6.1 for Eq. (6.22)), and Chap. 4 for Eq. (6.24).

Exercises

6.1 Referring to Video 6.1: Determine using the description of the magnetic potentiometer in Sect. 6.2

a) The current I in the vertical conductor in Fig. 6.8;

b) the average magnetic field H in the long coil of Fig. 6.11 (its data are given in the figure caption); and

c) the magnetic potential difference $\int H \mathrm{d}s$ of the permanent magnet in Fig. 6.12 (Sect. 6.3).

6.2 We imagine a bundle of identical long coils with square cross-sections as in Fig. 4.6. Their windings have N turns, their length is l, and all of them are carrying the same current I. One coil is pulled out of the interior of the bundle.

a) In the 'tunnel' which is left where the coil was removed, we insert a magnetic potentiometer. How large is the magnetic potential difference U_{mag} when the ends of the potentiometer are touching each other outside the bundle?

b) Compare the magnetic potential differences which would be measured if the potentiometer is either completely within the tunnel ($U_{\text{mag,i}}$), or, as shown in Fig. 6.11, it forms a curve completely outside the bundle ($U_{\text{mag,o}}$) (Sect. 6.3).

6.3 In Fig. 6.12, the magnetic potential $U_{\text{mag,a}}$ was measured outside a permanent magnet between its ends (Video 6.1). Now imagine that a thin channel has been bored through the magnet along its length. Compare the magnetic potential $U_{\mathrm{mag,i}}$ that would be measured along this channel between its ends, with $U_{\mathrm{mag,a}}$. (Sect. 6.3, 14.5)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_6) contains supplementary material, which is available to authorized users.

How the Fields Depend on the Frame of Reference

7

7.1 Preliminary Remarks

In this chapter, we want to show that electric and magnetic fields depend upon the *frame of reference* from which they are observed. This fact will lead us to a deeper understanding of induction in moving conductors, as already described in Chap. 5, Sect. 5.2 in the groups 2 and 3.

7.2 A Quantitative Evaluation of ROWLAND's Experiment

In ROWLAND's experiment (Sect. 4.3), the magnetic effects of a *conduction current* could be imitated by moving charges alone. But this is not the true significance of the experiment; that lies in a logical conclusion which can be drawn directly from it.

We return to Fig. 4.12 and imagine that a condenser consists of two rings which are rotating in the same direction around a common axis, so that they are moving with an orbital velocity $\boldsymbol{u}$. Then we observe the field-line pattern as sketched in Fig. 7.1. The magnetometer M is placed in this sketch at rest between the two plates, that is in the region of greatest strength of the *electric* field E.

As a result of the motion of the condenser, in addition to its electric field, a magnetic field is produced. Its field lines are perpendicular to those of the electric field and also perpendicular to the direction of the relative motion. This is a very significant fact; we now deal with its quantitative formulation.

If the ring-shaped charge carriers have a very large radius r, we can neglect their curvature over comparatively long regions and consider the direction of their velocity $\boldsymbol{u}$ to be constant. Between the plates, which have an area of $A = b\,l\;(l = 2\pi r)$, there is an electric displacement density $\boldsymbol{D}$ of magnitude $D = Q/A$ (see Sect. 2.13, Eqns. (2.4) and (2.5)). The rotating charged rings produce a magnetic field $\boldsymbol{H}$. It has a non-zero magnitude only between the two closely-spaced rings and is to a good approximation homogeneous there. On a closed path of length $2b$ which passes around one of the rings, the magnetic potential difference is $U_{\text{mag}} = H_z\,b$ (neglecting the field in the exterior

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_7

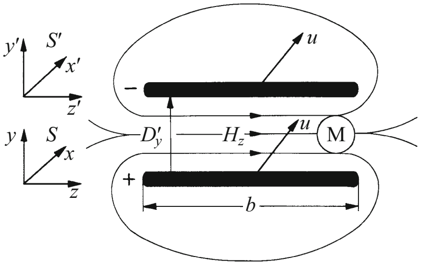


Figure 7.1 Magnetic field lines from positive and negative electric charges which are moving parallel to each other (perpendicular to and into the plane of the page) with their plate-shaped carriers, both at the same velocity $\boldsymbol{u}$. *Left*: Two right-handed coordinate systems. The upper system S' is at rest relative to the plates, while the lower system S (the 'laboratory system') is at rest relative to the room and the magnetometer M (a compass needle with a light pointer). S' is moving relative to S at the velocity $\boldsymbol{u}$. Above and below the pair of plates, the magnetic fields essentially cancel each other.

region). This is equal to the current $I = Qu/l$ enclosed by the path. Then the observer in S finds: $H_z b = Qu/l$, or

$$H_z = \frac{uQ}{lb} = u\frac{Q}{A} = u\,D_y\,,$$

and, in general form and vector notation,

$$\boldsymbol{H} = (\boldsymbol{u} \times \boldsymbol{D}) \quad \text{or} \quad \boldsymbol{B} = \varepsilon_0\mu_0(\boldsymbol{u} \times \boldsymbol{E})\,. \tag{7.1}$$

This field occurs in addition *to the electric field when the condenser, the carrier of the electric field, is* moving *relative to the observation apparatus* (the magnetometer M in Fig. 7.1) *with the velocity* $\boldsymbol{u}$. The definition of the vector $\boldsymbol{u}$ can be seen from Fig. 7.1.

In order to obtain Eq. (7.1) with the aid of the theory of relativity,C7.1 let us consider the condenser to be at rest in the system S' and the magnetometer to be at rest in the system S. S' is moving at the velocity $\boldsymbol{u}$ in S.C7.2 Then the theory states that:

$$\boldsymbol{B} = \gamma\varepsilon_0\mu_0(\boldsymbol{u} \times \boldsymbol{E}')\,, \tag{7.2}$$

$$\boldsymbol{E} = \gamma\boldsymbol{E}'\,, \quad \gamma = \frac{1}{\sqrt{1 - u^2/c^2}}\,, \tag{7.3}$$

where $c = 3 \cdot 10^8$ m/s is the velocity of light in vacuum. For $u \ll c$ (i.e. $\gamma \approx 1$), we obtain from Eq. (7.2) the equation (7.1), *which can thus be considered to be experimental evidence for the theory of relativity*.C7.3 The origin of the factor γ is the LORENTZ contraction, which makes the condenser, at rest in S', appear to be foreshortened along the direction of $\boldsymbol{u}$ in S (see also Sect. 7.4). This increases the charge density on its plates, and thereby also $\boldsymbol{E}$ and $\boldsymbol{B}$. This (small) effect was neglected in deriving Eq. (7.1).

C7.1. Introductions to the special theory of relativity can be found in many textbooks. See for example M. Born, "*Einstein's Theory of Relativity*" (Dover Publications, 1965) (available online for download at https://archive.org/details/einsteinstheoryo00born). Other introductions with examples include R.P. Feynman, *Lectures on Physics*, Vol. I, Chaps. 15–17 and 34, and Vol. II, Sect. 13.6 for a relativistic treatment of moving charges. This can be read online at http://www.feynmanlectures.caltech.edu/ See also "Electricity and Magnetism" by E.M. Purcell and D.J Morin, 3rd ed., Cambridge University Press (2013), Sec. 5.9.

C7.2. See Fig. 5.4 in Sect. 5.2.

C7.3. The principle of relativity requires that even an observer in S', where the charges are at rest, so that only an electric field, but no magnetic field is present, would find that the magnetometer (compass needle) is subject to a torque. This is certainly the case, but the demonstration that a magnetometer which is moving relative to an electric field experiences a torque, analogous to the LORENTZ force (Sect. 7.3), was too difficult even for a scientist like HENRY ROWLAND (1848–1901)!

7.3 Induction in Moving Conductors

The earlier experiments (Figs. 5.10 and 5.11) are summarized in a schematic sketch in Fig. 7.2. There, the points where the magnetic field lines of the homogeneous field perpendicular to the paper pass through the plane of the page are shown as black dots. The slider of length L is moving relative to the frame of the magnetic field (the field coil) at the velocity $\boldsymbol{u}$ perpendicular to the direction of the magnetic field lines. The observer finds negative charges at the point C, and positive charges at A; and, using the arrangement shown in Fig. 5.10, the magnitude of the induced voltage is found to be:

$$U = BuL\,. \tag{5.9}$$

This experiment can demonstrate the validity of the principle of relativity convincingly, by moving either the conductor (slider) or the field coil which produces the magnetic field. In both cases, the same induced voltage is observed. The interpretation of this result depends on the frame of reference of the observer.

Initially, the observer is at rest next to the field coil in the reference frame S, and describes the induction effect as follows:

'Like every object, the slider contains electric charges, an equal number of each sign. These charges participate in the motion of the slider at the velocity $\boldsymbol{u}$. During the motion, they collect at the points A and C (Fig. 7.2); the slider can serve as a *current source*. Therefore, within it, there must be *forces* $\boldsymbol{F}$ *which separate the charges*, as in any current source. Here, they are a result of the motion. They pull positive charges downwards and negative charges upwards'.

While these forces $\boldsymbol{F}$ are piling up charges at the points A and C, according to Eq. (5.9), an electric voltage is produced between A and C. It, in turn, causes a restoring force $\boldsymbol{F}^*$ to act on the two separated charges (Eq. (3.1)). In vector form, this restoring force is given by:

$$\boldsymbol{F}^* = Q(\boldsymbol{B} \times \boldsymbol{u})\,. \tag{7.4}$$

For positive charges, it acts upwards, and for negative charges it acts downwards. If only these forces were acting, the charges would again be pulled together. The experimentally-observed stationary state is

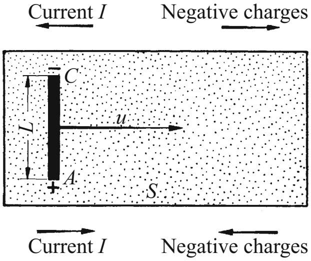


Figure 7.2 Schematic drawing of induction in moving conductors. The magnetic field points perpendicular to the plane of the page towards the reader. The circular cross-section of the field coil seen in Fig. 5.10 has been replaced here by a rectangular cross-section. The rod A-C corresponds to the slider in Fig. 5.10.

therefore possible only if the forces $\boldsymbol{F}$ that separate the charges are equal and opposite to the forces $\boldsymbol{F}^*$ in Eq. (7.4), i.e. $\boldsymbol{F} = -\boldsymbol{F}^*$. Then for the *forces* $\boldsymbol{F}$, we have

$$\boldsymbol{F} = Q(\boldsymbol{u} \times \boldsymbol{B})\,. \tag{7.5}$$

These forces, named for H.A. LORENTZ, act perpendicular to both $\boldsymbol{u}$ and $\boldsymbol{B}$ on the charges Q which are moving at the velocity $\boldsymbol{u}$ in the magnetic field $\boldsymbol{B}$. For the observer who is at rest in the frame of the field coil, the LORENTZ force is a new experimental fact.

Now, we let the observer move with the slider (in the frame of reference which we have called S'). For this observer, the charges are at rest. Therefore, the magnetic field with its LORENTZ force *cannot* be acting on them, but instead, only an *electric* field. This field, as shown by the experiment, is given by[C7.4]

C7.4. All of the quantities measured in the system S', insofar as they are different from those measured in S, will likewise be written with a prime.

$$\boldsymbol{E}' = \frac{\boldsymbol{F}'}{Q} = \boldsymbol{u} \times \boldsymbol{B}\,. \tag{7.6}$$

From this point of view, an electric field occurs in S' in addition *to the magnetic field*, when the field coil, the source of the magnetic field, is *moving at the velocity* $-\boldsymbol{u}$ *relative to the observation apparatus* (the slider in Fig. 7.2). Induction in moving conductors is therefore a counterpart to ROWLAND's experiment – only the roles of the electric and the magnetic fields have been reversed.

An electric field which appears during the relative motion in the frame of reference S' would be observable in principle without a conducting slider. Imagine that the slider is a rod made of a heated artificial resin. During the motion, it would cool and its interior state would be "frozen in". When it was removed from the field, the rod would act as an electret (Sect. 3.10, Point 2b), with negative charges at the top end and positive charges at the bottom.

Wilhelm WIEN passed fluorescent molecules at a high velocity through a homogeneous *magnetic* field instead of moving a slider rod. The *electric* field which then acts on the electrons in the molecules was detected via the STARK effect, which is a splitting of the spectral lines into several components in an electric field (13th edition of "*Optik und Atomphysik*", Chap. 14, Sect. 47; or see e.g. hyperphysics.phy-astr.gsu.edu/hbase/atomic/stark.html).

7.4 Fields and the Principle of Relativity

In Chap. 5, we had to distinguish *experimentally* between two types of induction, namely induction in a coil at rest (Sect. 5.3), and induction in moving conductors (Sect. 5.5).

In the former case, the explanation was as follows: With induction in a coil at rest, the forces that separate the charges are produced by an *electric* field. This field occurs *in addition* to the magnetic field, *dur-*

ing the *variation* of the magnetic field. It surrounds the magnetic field with closed, circular electric field lines (Sect. 6.1, Fig. 6.2; the mathematical formulation is given as the MAXWELL equation (6.22)).

According to Sect. 7.3 also, in the second case, that is induction in moving conductors, the forces which separate the charges are due to an *electric* field (which is seen by an observer who is at rest relative to the moving conductor). However, it remained completely unexplained just how the motion was able to produce an electric field. The answer is given only by the theory of relativity:[C7.1] It states that the windings of the field coil are no longer exactly electrically neutral during its motion. They acquire an excess of charges of each sign, and between the excess charges, an electric field acts. We can already show this here; from the theory of relativity, we need only the LORENTZ contraction: An observer whose frame of reference (the S system) is at rest relative to some object measures its length to be l. An observer who is moving parallel to this length relative to the object with a velocity u measures in his/her frame of reference S' a reduced ('contracted') length

$$l' = l\sqrt{1 - u^2/c^2} \tag{7.7}$$

(c = light velocity in vacuum = $3 \cdot 10^8$ m/s).

Following this brief but sufficient introduction to the LORENTZ contraction, we repeat in Fig. 7.3 the content of Fig. 7.2. The terminology needed is given in the caption of this figure. – Initially, an observer is at rest relative to a *field coil* through which the current I is flowing (S system). In the windings of this coil, the positive charges (the lattice ions) are fixed, and the negative charges (the electrons) can move at a rather small velocity[1] u_e, i.e. we have $u_e \ll c$. The magnitudes of the charges q and their densities are equal within the windings of the field coil; the conductor is electrically neutral, so that

$$\varrho_+ = \varrho_- = \frac{q}{V} = \varrho \tag{7.8}$$

($V = ld^2$ is the volume of a wire of length l).

For an observer at rest relative to the *slider* (S' system), the positive charges which are fixed within the wires are moving at the velocity $-u$ (Fig. 7.3); for this observer, in the windings of the field coil just below (3) and just above (1) in the figure, the positive charges have the *same* charge density

$$\varrho'_+ = \frac{q}{d^2 l'} = \frac{q}{d^2 l\sqrt{1 - u^2/c^2}} = \frac{\varrho}{\sqrt{1 - u^2/c^2}}\,,$$

or, for $u \ll c$ (i.e. $\gamma \approx (1 + u^2/2c^2)$)

$$\varrho'_+ = \varrho\left(1 + \frac{1}{2}\frac{u^2}{c^2}\right)\,. \tag{7.9}$$

[1] Compare the footnote in Sect. 8.3.

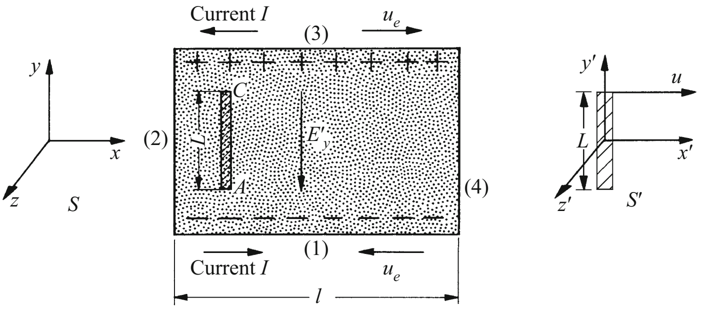


Figure 7.3 Induction in a moving conductor. The sketch shows the rectangular cross-section of a long field coil, whose long axis is perpendicular to the plane of the page. It is at rest in the frame of reference S, whose z axis is also perpendicular to the plane of the page. The coil has along its length l^* all together N turns in its windings, made of wire with a square cross-section of area d^2. When a current I is passed through this coil, it produces a homogeneous magnetic field of magnitude $H = NI/l^*$. This field is perpendicular to the plane of the page and points towards the reader, and the passage of its field lines through the page is marked by dots. The slider (of length L) is at rest in the frame of reference S', which is moving at the velocity $\boldsymbol{u}$ parallel to the x axis (as is sketched to the right in the figure for clarity). The length l, the current I and the velocity of the electrons (drift velocity) u_e are all measured in S. The excess charge densities (+ and −) are observed in S', as is the electric field $E'_{y'}$. As observed in S', the system S is moving to the left (at the velocity $-u$).C7.5 (**Exercise 5.6**)

C7.5. The goal of the following calculations is to derive the electric field (in the form $\boldsymbol{E}' = \boldsymbol{D}'/\varepsilon_0$) which is observed in the moving system S', moving with the velocity u of the slider. The field is shown in Eq. (7.14) to be
$E' = \frac{\gamma}{\varepsilon_0} \frac{Nq}{l\,l*} \cdot \frac{u_e\,u}{c^2}$.
This field is caused by the relative motion of the slider and the magnetic field source. It is a relativistic effect and does not appear in the laboratory frame S, where the field source is at rest. As seen in Eq (7.16), the field E' is proportional to u and B.

For the same observer in the S' system, the negative charges (electrons) in the upper section of the windings (3) have a velocity $(u - u_e)$, and in the lower section (1) a slightly greater velocity $(u + u_e)$. This observer thus finds the charge density of the *negative* charges in the S' system (with $u_e \ll u$) in the *upper* section of windings just below (3), to be

$$\varrho'_- = \frac{\varrho}{\sqrt{1-(u-u_e)^2/c^2}} = \varrho\left(1 + \frac{1}{2c^2}(u^2 - 2uu_e)\right), \qquad (7.10)$$

and for the *lower* section of windings, just above (1),

$$\varrho'_- = \frac{\varrho}{\sqrt{1-(u+u_e)^2/c^2}} = \varrho\left(1 + \frac{1}{2c^2}(u^2 + 2uu_e)\right). \qquad (7.11)$$

For the upper section of the coil containing N wires (3), combining Eqns. (7.10) and (7.9) with Eq. (7.8) yields an *excess of positive charges*:

$$\Delta Q'_+ = N\Delta q'_+ = \frac{Nqu_e u}{c^2}. \qquad (7.12)$$

For the lower section of windings (1), the combination of Eqns. (7.11) and (7.9) with Eq. (7.8) yields an *excess of negative charges*:

$$\Delta Q'_- = N\Delta q'_- = -\frac{Nqu_e u}{c^2}. \qquad (7.13)$$

Therefore, between $\Delta Q'_+$ and $\Delta Q'_-$, there is an electric field which is produced by the LORENTZ *contraction*. Dividing $\Delta Q'_+$ by the cross-sectional area $l'l^*$ of the windings gives its displacement density $D' = \varepsilon_0 E'$, and thus the electric field:[C7.6]

$$E' = \frac{1}{\varepsilon_0}\frac{Nqu_\mathrm{e}u}{l'l^*c^2} = \frac{1}{\varepsilon_0}\frac{Nqu_\mathrm{e}u}{ll^*c^2\sqrt{1-u^2/c^2}} = \frac{\gamma}{\varepsilon_0}\frac{Nqu_\mathrm{e}u}{ll^*c^2} \qquad (7.14)$$

(γ was defined in Sect. 7.2).

We know that $qu_\mathrm{e} = Il$ and $\mu_0 NI/l^* = B$. Inserting these quantities into (7.14) gives

$$E' = \frac{\gamma\, u}{\varepsilon_0\mu_0 c^2}B\,. \qquad (7.15)$$

Applying the relation $1/c^2 = \varepsilon_0\mu_0$ (see Sect. 8.2) and taking the directions into account, we finally obtain the electric field in vector notation:[C7.7]

$$\boldsymbol{E}' = \gamma(\boldsymbol{u}\times\boldsymbol{B})\,, \qquad (7.16)$$

and for $u \ll c$, Eq. (7.6) follows from this. The field $\boldsymbol{E}'$ is directed downwards in Fig. 7.3: The electric field which results from the motion of the field coil relative to the slider in S' thus has the magnitude and direction that were determined empirically in Sect. 7.3. Result: The observer who is at rest relative to the slider can explain the observed charge separation in terms of the LORENTZ contraction.[C7.8]

Now, finally, we are in a position to understand the experiments of the second group in Sect. 5.2 in detail: As seen from the reference frame S of the field coil, the LORENTZ *force* produces the induced voltage in the induction coil, which is at rest in S'. As seen from the induction coil in the reference frame S', the LORENTZ *contraction* of the charge densities in the field coil produce the same voltage in the induction coil. In the general case of induction in moving conductors, one must of course also take into account (in addition to the LORENTZ force or the LORENTZ contraction) the possible time dependencies of the magnetic field, as described in Sect. 5.6.

7.5 Summary: The Electromagnetic Field

Because of their importance, the results of this chapter will be briefly summarized here:

1. An observer can place a measuring apparatus (the magnetometer in Fig. 7.1 or the moving conductor in Fig. 7.3) at rest within a frame of reference which is moving at a velocity u relative to the source of the field (condenser or field coil). Then she or he will observe

[C7.6] C7.6. The lengths l^* and d have no components along the direction of $\boldsymbol{u}$, so they are not affected by the LORENTZ contraction.

[C7.7] C7.7. This field can also be derived directly from the theory of relativity. For this derivation, we consider that the field coil is at rest in the frame S (Fig. 5.4 in the footnote in Sect. 5.2) and the charge Q is at rest in S'. Then we have

$$\boldsymbol{E}' = \gamma(\boldsymbol{u}\times\boldsymbol{B})\,,$$

and

$$\boldsymbol{B}' = \gamma\boldsymbol{B}\,.$$

In S', then, the force $\boldsymbol{F}' = Q\gamma(\boldsymbol{u}\times\boldsymbol{B})$ will be observed. Using a relativistic transformation, we can compute from it the force $\boldsymbol{F}$ which will be observed in S. The transformation equation for forces in this case (where the geometry is particularly simple) gives

$$\boldsymbol{F} = \frac{1}{\gamma}\boldsymbol{F}' = Q(\boldsymbol{u}\times\boldsymbol{B})\,.$$

This is precisely the LORENTZ force! It thus follows directly from the theory of relativity. (Note again: The field $\boldsymbol{E}$ measured in S is zero.) More details may be found for example in P. Lorrain, D.R. Corson, and F. Lorrain, "*Electromagnetic Phenomena*", Freeman, New York (2000), Chap. 13.

[C7.8] C7.8. The explanation given here can also be applied to the experiment shown in Fig. 5.11. The reader should also try to apply it to BARLOW's wheel!

in addition to the electric field also a magnetic field (Fig. 7.1) or *in addition* to the magnetic field also an electric field (Fig. 7.3). In the first case, Eq. (7.1) applies, and in the second case, Eq. (7.6). For $u = 0$, the additional fields vanish. They are thus purely *relativistic* effects.

2. Electric fields and magnetic fields are not autonomous and independent of one another. They are both parts of an *electromagnetic field*. Their presence or absence depends upon the frame of reference chosen for the observations.[C7.9]

C7.9. Making use of this newly-found information, the reader should reconsider the remarks made at the end of Chap. 5 following Eq. (5.11)!

3. Induction in moving conductors can occur as a result of the LORENTZ contraction. Or, conversely: *Induction in moving conductors is an experimental verification of the existence of the* LORENTZ *contraction*.

Forces in Magnetic Fields 8

8.1 Demonstration of the Forces on Moving Charges

From our detailed consideration of induction in moving conductors, we became aware of the existence of the LORENTZ force

$$\boldsymbol{F} = Q(\boldsymbol{u} \times \boldsymbol{B}) \tag{7.5}$$

(F for example in newton ($= 1\,\mathrm{N}$), B in $\mathrm{V\,s/m^2}$ ($= 1\,\mathrm{T}$), u in m/s, Q in A s).

This force acts on a charge Q which is moving at the velocity $\boldsymbol{u}$ in a magnetic field of flux density $\boldsymbol{B}$. As can be seen from the vector product, the force is perpendicular both to the magnetic field and to the velocity of the charge (Fig. 8.1):

Unfortunately, we cannot verify this equation (7.5) with a demonstration experiment using a mechanically-moved macroscopic charge carrier, for example a charged soap bubble. We cannot make the product Qu sufficiently large for such a large carrier of charge.[C8.1] However, we can test Eq. (7.5) and its consequences as shown in Fig. 8.1 experimentally in other ways.

In Sect. 4.3 (the ROWLAND experiment), the macroscopic motion of a charge carrier was found to be equivalent to the invisible motion of

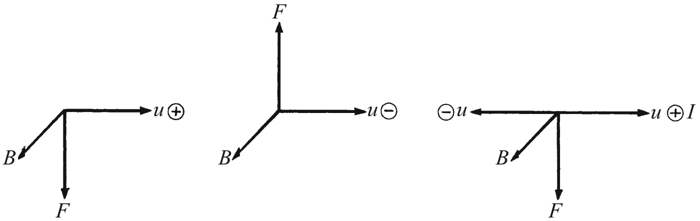
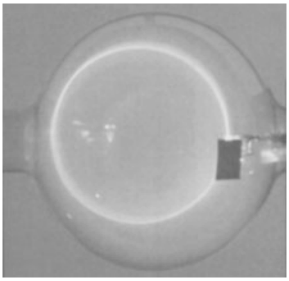


Figure 8.1 Forces $\boldsymbol{F} = Q(\boldsymbol{u} \times \boldsymbol{B})$ on moving charges. They are perpendicular to both the velocity $\boldsymbol{u}$ and to the magnetic flux density (also known as the magnetic induction field) $\boldsymbol{B}$. In the two left-hand images, the charges have opposite signs but the same direction of motion (as e.g. in Figs. 5.10 to 5.12b). In the right-hand image, they have opposite directions of motion. In the general case, $\boldsymbol{u}$ and $\boldsymbol{B}$ are not perpendicular to each other.

C8.1. This difficulty is avoided by using an electron beam, as shown here in the figure (Photo: K. Lechner, IWF Göttingen):



In a glass bulb containing hydrogen gas at a pressure of $p = 1$ Pa (from the Leybold Co., Cologne), an electron beam is moving through a magnetic field that points perpendicular to the plane of the page. The beam is accelerated by a voltage of ca. 200 V, so that its orbital velocity u is ca. 10^7 m/s. The field is produced by a pair of HELMHOLTZ coils (B ca. $8 \cdot 10^{-4}$ T, directed towards the reader). The beam is deflected into a circular orbit by the field (Lorentz force), with a counter-clockwise sense of rotation and a radius $r = 6 \cdot 10^{-2}$ m. The LORENTZ force is thus constant at every point along the orbit and is perpendicular to the momentary orbital velocity, acting towards the center of the circular orbit.

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_8

The frequency of rotation of the electrons, the so-called cyclotron frequency, is given by $\nu = QB/2\pi m$ ($m = 9.11 \cdot 10^{-31}$ kg is the mass of the electron), and is independent of the (non-relativistic!) velocity u (See **Exercise 8.1**).

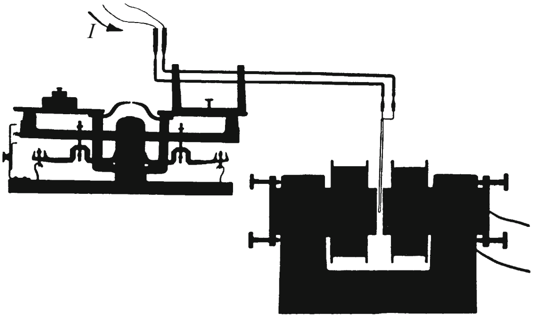
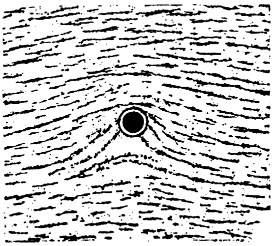


Figure 8.2 A horizontal current-carrying conductor perpendicular to the homogeneous magnetic field of an electromagnet. The conductor appears foreshortened due to the perspective. Numerical example: $I = 15\,\mathrm{A}$, $l = 5 \cdot 10^{-2}\,\mathrm{m}$, $B = 1.5\,\mathrm{V\,s/m^2}$, $F = 1.5\,\mathrm{V\,s/m^2} \cdot 15\,\mathrm{A} \cdot 5 \cdot 10^{-2}\,\mathrm{m} = 1.13\,\mathrm{W\,s/m} = 1.13\,\mathrm{N\,m/m} = 1.13\,\mathrm{N}$.

Figure 8.3 The field-line pattern corresponding to Fig. 8.2. The conductor is perpendicular to the plane of the page.C8.2



C8.2. Compare the field-line patterns in Fig. 8.3 with the patterns of the electric field lines in Fig. 3.9. In perfect analogy to the electric case, the force is determined by the magnetic field of one current distribution (here: of the electromagnet) and the current in another conductor (see Eq. (8.1)).

electric charges within a conductor. The observer can use the charge carriers of either sign, Q or $-Q$, within a conductor of length l as a ‘rest’ frame of reference. In that frame, only the other charges have a velocity u. Quantitatively, we found

$$Qu = I\,l\,. \tag{4.4}$$

This equation for the magnitudes can be inserted into Eq. (7.5), leading to the force on a section of conductor of length l which carries a current I perpendicular to the field B:

$$F = IB\,l\,. \tag{8.1}$$

To verify this equation, we make use of a horizontal, linear conductor in a homogeneous magnetic field as in Fig. 8.2. The field B is produced by an electromagnet and is directed perpendicular to the conductor. The conductor forms a trapeze with its two rigid lead wires and is hanging from a force meter (balance). A numerical example is given in the legend of the figure. The pattern of field lines is shown in Fig. 8.3.

8.2 Forces Between Two Parallel Currents

As an application of Eq. (8.1), we compute the forces between two parallel, straight conductors of length l at a distance r from each other, which are carrying the currents I_1 and I_2 (Fig. 1.9). The current I_1 produces

$$\text{the field strength} \quad H = \frac{1}{2\pi}\frac{I_1}{r} \tag{6.13}$$

at a distance r, that is

$$\text{the flux density} \quad B = \frac{\mu_0}{2\pi}\frac{I_1}{r}\,. \tag{8.2}$$

Equations (8.1) and (8.2) together yield the attractive force (when the currents in both conductors are flowing in the same direction) or the repulsive force (when they are opposite):

$$F = \frac{\mu_0}{2\pi}\frac{I_1 I_2 l}{r}\,. \tag{8.3}$$

Numerical example: $I = 100\,\text{A}$, $l = 0.5\,\text{m}$, $r = 1\,\text{cm}$, $\mu_0 = 4\pi \cdot 10^{-7}\,\text{V s/A m}$, $F = 10^{-1}\,\text{N}$.

We now apply Eq. (8.3) to a *special case*: We imagine that the two currents are carried by two identical strings of charges which flow beside each other through space (Fig. 8.4) (these could be electrically-charged particle beams with a linear charge density Q/l). *Thus, in contrast to the usual conductivity currents in metals, etc., the 'background charges' of opposite sign are missing here*. As a result, in addition to the attractive magnetic force F_{magn} between the two strings, there is a repulsive electrical force F_{el} perpendicular to the direction in which the charges are moving.

For the *magnetic attractive force*, we obtain from the combination of Eqns. (8.3) and (4.4):

$$F_{\text{magn}} = \frac{\mu_0}{2\pi}\frac{Q^2 u^2}{rl}\,. \tag{8.4}$$

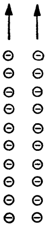


Figure 8.4 Two parallel strings of charges of the same sign, moving in the same direction

For the electrical repulsion, we find

$$F_{\text{el}} = \frac{1}{2\pi\varepsilon_0}\frac{Q^2}{rl}\,. \tag{8.5}$$

Derivation: The charge string at the left produces an electric field strength at a distance r on a cylindrical surface of $E = \frac{1}{2\pi\varepsilon_0}\frac{Q}{rl}$. This acts according to Eq. (3.5) on the charge string at the right, exerting a force

$$F_{\text{el}} = QE = \frac{1}{2\pi\varepsilon_0}\frac{Q^2}{rl}\,.$$

Taking the ratio of Eqns. (8.4) and (8.5) yields

$$\frac{F_{\text{magn}}}{F_{\text{el}}} = \varepsilon_0\mu_0 u^2\,. \tag{8.6}$$

In this equation, the product $\varepsilon_0\mu_0 u^2$ must be a pure number; it represents the ratio of two forces. Therefore, $1/\sqrt{\varepsilon_0\,\mu_0}$ must refer to a velocity. A calculation gives

$$\frac{1}{\sqrt{8.859\cdot 10^{-12}\,\frac{\text{A s}}{\text{V m}}\cdot 4\pi\cdot 10^{-7}\,\frac{\text{V s}}{\text{A m}}}} = 2.998\cdot 10^8\,\frac{\text{m}}{\text{s}}\,.$$

This velocity[1] is exactly the same as the velocity of light c in vacuum! We thus have the experimental result that

$$\frac{1}{\sqrt{\varepsilon_0\mu_0}} = c\,. \tag{8.7}$$

This is not simply an accidental agreement, but rather a fundamental relation between the velocity of light and electromagnetic phenomena. The first to recognize this fact in its wide-ranging consequences was James Clerk MAXWELL, in 1862: He explained light waves as short-wavelength *electromagnetic* waves (at the time experimentally still unobserved!).

Inserting Eq. (8.7) into Eq. (8.6), we obtain

$$F_{\text{magn}} = F_{\text{el}}\frac{u^2}{c^2}\,. \tag{8.8}$$

In words: Under similar geometric conditions, the *magnetic* forces produced by moving electric charges are smaller by the factor u^2/c^2 than the *electric* forces between the same charges at rest. We want to elucidate this statement by referring to Fig. 8.5:

[1] W. WEBER and R. KOHLRAUSCH in 1856 described this velocity simply as a "critical" value which could make magnetic forces just as strong as electrical forces.

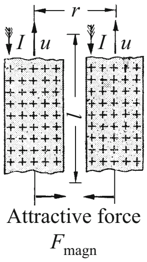


Figure 8.5 A numerical example to elucidate Eq. (8.8): Two parallel copper wires of length $l = 1\,\text{m}$ and $1\,\text{mm}^2$ cross-sectional area are a distance $r = 0.1\,\text{m}$ apart, and carry a current of $I = 6\,\text{A}$. According to the footnote below, the current I is due to negative charges of $Q = 1.36 \cdot 10^4\,\text{A s}$ which are moving at a velocity $u = 0.44\,\text{mm/s}$. We thus find $(u/c)^2 = 2.16 \cdot 10^{-24}$. The current flowing on the right produces a magnetic flux density (induction field) of $B = 1.2 \cdot 10^{-5}\,\text{V s/m}^2$ at the position of the left wire (Eq. 8.2). This causes a force $F_{\text{magn}} = QuB = 7.23 \cdot 10^{-5}\,\text{N}$ to act on the charge Q there. From Eq. (8.5), the mobile negative charges Q *alone* (i.e. without the presence of the equally strong positive background charges) would repel each other with a force $F_{\text{el}} = 3.34 \cdot 10^{19}\,\text{N}$. Then the ratio of forces is $F_{\text{magn}}/F_{\text{el}} = 2.16 \cdot 10^{-24} = (u/c)^2$.

Fig. 8.5 shows schematically two current-carrying wires (not to scale!). The mobile negative charges (electrons) within them move as gray clouds through the lattice of fixed positive charges (lattice ions), at velocities which are normally less than 0.5 mm/s.[2] Thus, $(u/c)^2$ is only a tiny fraction, on the order of 10^{-24}! If the positive ions were lacking, the two clouds of negative charge would repel each other with the force F_{el}. The magnetic field generated by their motion produces an attractive force between them of only $F_{\text{magn}} \approx 10^{-24}\,F_{\text{el}}$.

For this reason, Eq. (8.8) leads to the following conclusion: The production of magnetic forces through electric currents is certainly of

[2] The copper wires used in everyday house wiring are normally limited to current densities of less than $6\,\text{A/mm}^2$. A copper wire with a cross-sectional area of $1\,\text{mm}^2$ and a length of 1 m has a mass $m = 8.95$ g, and thus the amount of substance $n = 0.14\,\text{mol}$ (Vol. 1, Sect. 13.1). It contains copper ions, each carrying one positive electrical elementary charge. For these ions, the FARADAY constant is the quotient

$$\frac{\text{Charge } Q}{\text{Amount of substance } n} = 9.65 \cdot 10^4\,\frac{\text{A s}}{\text{mol}}\,.$$

The copper wire therefore contains a positive ionic charge of $Q = 1.36 \cdot 10^4\,\text{A s}$. The mobile negative charge Q between the lattice ions is just as large. Inserting this quantity of charge into Eq. (4.4) gives

$$u = \frac{I\,l}{Q} = \frac{6\,\text{A} \cdot 1\,\text{m}}{1.36 \cdot 10^4\,\text{A s}} = 4.4 \cdot 10^{-4}\,\frac{\text{m}}{\text{s}} = 0.44\,\frac{\text{mm}}{\text{s}}\,.$$

eminent importance for electrical technology; but physically, it belongs among the “*second-order effects*” or “*relativistic effects*”.[C8.3]

C8.3. The derivation of this equation based on the theory of relativity can be found in “Electricity and Magnetism” by E.M. Purcell and D.J. Morin, 3rd edition, Cambridge University Press (2013), Sect. 5.9, p. 264. Here, instead of two rows of point charges, the repulsive electric force and the attractive magnetic force between two point charges (protons) moving with the same velocity $\boldsymbol{v}$ and separated by the distance r have been derived (in Eq. (5.29) there; the first term on the left and the term on the right). Their ratio yields Eq. (8.8).

8.3 LENZ’s Law and Eddy Currents

Induction processes give rise to electric fields, currents and forces. The sign of their direction is determined from the rule formulated by H. F. E. LENZ (1834):

The electric fields, currents and forces produced by induction *are always directed to oppose the processes which give rise to that induction.* This results from the conservation of energy. If the sign were opposite, the process causing the induction would be subject to a positive feedback and would increase without limit, creating energy from nothing. Examples (see also **Video 8.1**):

Video 8.1:
“LENZ’s Law”
http://tiny.cc/mcggoy
The video shows an impressive demonstration experiment of LENZ’s law from the collection of physics demonstrations at Cornell University. (Note as well the *wobble oscillations* exhibited by the aluminum ring at the end of the video (see also Video 11.12 in Vol. 1, http://tiny.cc/qcgvjy).)

1. In Fig. 5.5, the induction was caused by the increase of the magnetic field from the field coil. According to LENZ’s law, the current induced in the induction coil must impede this increase of the magnetic field. It must therefore flow in the opposite direction to the current in the field coil.

2. In Fig. 8.6, an aluminum ring as “induction coil” is hanging between the conical pole pieces of a horseshoe magnet. We pull the magnet away on its guide rail. The ring follows it. Their separation, which is the cause of the induction, is opposed.

3. Now we reverse the experiment, i.e., we push the magnet towards the ring and try to force the ring into the region between the magnet pole pieces where the field is strongest. The ring moves away from the approaching magnet; its approach, which is the cause of the induction, is again avoided.

4. In cases 2 and 3, we could make the hole in the ring as small as we want; finally, it becomes just a disk of metal. The currents which are induced in that disk are called *eddy currents*.

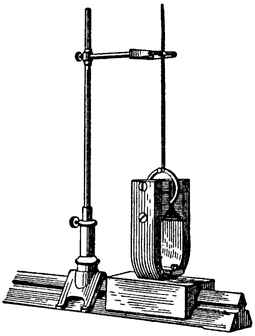


Figure 8.6 A ring-shaped “induction coil” is hanging like a pendulum between the poles of a horseshoe magnet which can be slid along a guide rail (an optical bench)

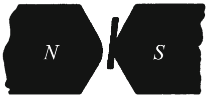


Figure 8.7 Eddy currents brake the fall of a silver coin in an inhomogeneous magnetic field **(Video 8.2)**

Video 8.2:
"The Eddy-Current Brake"
http://tiny.cc/1bggoy
A round aluminum disk is released as close as possible to the center in the region of strongest field (and therefore also the greatest field gradient). It has been previously cooled to the temperature of liquid nitrogen (77 K), which strongly enhances its electrical conductivity and therefore the braking effect. Then, a long aluminum plate is pulled through the field to demonstrate the astonishingly strong forces, which increase with the velocity of the plate.

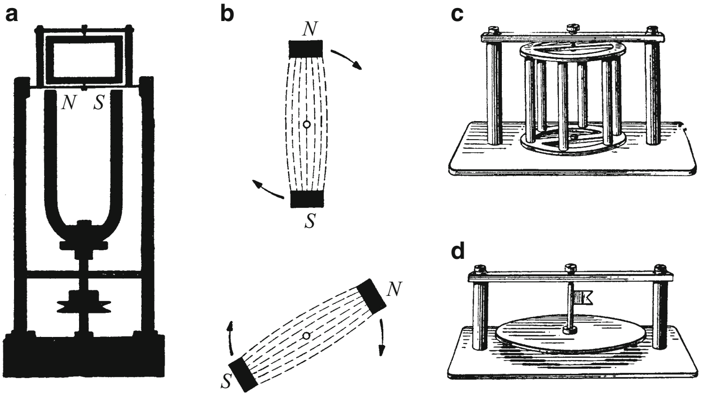


Figure 8.8 A rotating magnetic field with various "induction rotors". *Part b* shows a schematic drawing of the rotating magnetic field produced by the apparatus in a, at two different positions separated by 60°. The small circles mark the axis of rotation for an observer who is looking directly down from above. The magnetic field lines between its rotating poles *S*-*N* are indicated. *Parts c and d* show two rotors which can be used above the rotating magnet in place of the rectangular frame. The application of rotor *d* is a reversal of the experiment shown in Fig. 8.9. **(Video 8.3)**

Video 8.3:
"Induction Rotor"
http://tiny.cc/5bggoy
The principle of an induction motor is demonstrated with a setup as shown in a shadow projection in Fig. 8.8 a.

If we drop a silver coin through the *inhomogeneous* magnetic field of a large electromagnet (Fig. 8.7), then it doesn't fall freely with the usual acceleration in the air. Instead, it sinks slowly as if it were in a sticky fluid. Here, again, the induction process impedes its origin, in this case the falling motion of the coin.[C8.4]

C8.4. Modern coins which are made of alloys have lower electrical conductivities. This makes the induced currents smaller, and thus also the forces which slow the coin's fall.

5. We replace the linear motion by a rotation; in Fig. 8.8, we rotate a horseshoe magnet around its long axis and thereby obtain a "rotating magnetic field". Into this *rotating field* we then bring an "induction coil" which is mounted on bearings so that it can also rotate; it takes the simple form of a rectangular metal frame. The frame follows the rotation of the field. The angular motion between the field and the frame, which is the origin of the induction in the frame, is impeded. Soon, the frame is rotating almost as fast as the magnet. It cannot move *exactly* as fast; in that case, the field change within the area of the frame would no longer be present, so there would be no induction at all. The fractional velocity difference between the coil and the rotating field is called the *slip*. For the technological application of this experiment, the simple rectangular frame is replaced by a metal cage (Fig. 8.8c); one then refers to it as an induction rotor or 'shorted rotor' (see Fig. 9.21).

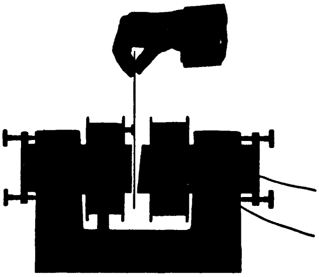


Figure 8.9 Eddy currents slow the rotation of a circular disk made of aluminum. Its axle lies far behind the plane of the page. The front surfaces of the magnet poles could also be parallel to each other, but then only the inhomogeneous regions at their edges would play a role in the braking process. (See also **Video 8.2**)

6. In the fourth experiment, we encountered *eddy currents*. They were produced by moving an *inhomogeneous* magnetic field through a metal plate of limited size. The magnetic flux passing through the plate changed with time.

Eddy currents can however also be produced without any changes in the geometrical arrangement of the components. In Fig. 8.9, we see a circular aluminum disk which extends into the *inhomogeneous* magnetic field of an electromagnet. The axle of the disk is well behind the plane of the drawing. The disk is hard to rotate; it shows a tough resistance to motion which is surprisingly strong. Induction of eddy currents again opposes its cause, the rotation of the disk.

The production of these eddy currents is best understood in terms of induction in moving conductors. In Fig. 8.10, we have sketched a cross section of the magnetic field and a segment of the disk. The small, dashed circle indicates a closed path for the electrons within the metal disk. All the electrons participate in the rotation of the disk. Therefore, they are moving perpendicular to the field lines, and LORENTZ forces $\boldsymbol{F} = Q(\boldsymbol{u} \times \boldsymbol{B})$ (Sect. 8.1) act, as indicated by the arrows. The flux density $\boldsymbol{B}$ of the field is greater below than above in the figure. F_3 is thus larger than F_1, and this gives rise to a circular motion of the electrons in a counter-clockwise direction. Furthermore, the forces F_2 shift the electron orbits to the right. These two motions are superposed and result in a cycloidal motion for the orbits of the eddy currents.

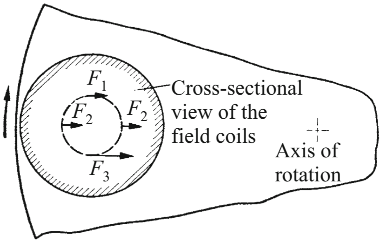


Figure 8.10 The production of eddy currents in the moving disk of Fig. 8.9 (***B*** is perpendicular to the plane of the page and points along the line of sight of the reader)

8.4 Damping of Rotating-Coil Instruments. The Creeping Galvanometer. Magnetic Flux Circuits

We refer to the second experiment in the previous section. There, we saw in Fig. 8.6 a metal ring mounted as a pendulum in a magnetic field. When it was set into swinging motion, it was seen to come to rest after only a few swings – it was strongly damped. The forces due to induction impede the oscillations of the pendulum (LENZ's law). This *induction damping* is often used to prevent unwanted vibrations. It is also called "eddy-current damping" or "eddy-current braking". Imagine the ring in Fig. 8.6 to be replaced by a metal disk.

Induction damping is in particular indispensable in the construction of various kinds of measurement instruments (including modern electronic analytical balances). It is used to prevent the annoying and time-consuming oscillations of the pointer before it settles down to its final position. As a rule, the *aperiodic limit*[3] can be reached, so that the pointer comes to rest in its final position after a minimal time.

We offer as one example the *induction damping of a rotating-coil ammeter* (Fig. 1.19). As a rule, it consists of two parts: First, one uses a rectangular metal frame as support for the coil windings. It acts like the ring in Fig. 8.6, and analogously for rotational oscillations. Second, the rotating coil itself can act as a closed-loop induction coil. The instrument is part of an electrical circuit; thus the ends of the rotating coil can be connected in some manner by a conducting link ("short circuit" or "shunt") as needed. The resistance of this shunt (for example in Fig. 1.33, around $10^6\ \Omega$) is called the "external resistance". A suitable choice of its magnitude allows experienced observers to attain the aperiodic limit for the pointer oscillations.

When the damping is too strong, the pointer *creeps* to its equilibrium position, without oscillating, but very slowly. This makes the instrument useless for measurements of the momentary values of currents or voltages. However, in contrast, a *creeping galvanometer* is very useful for the measurement of "current impulses" ($\int I\,\mathrm{d}t$) and "voltage impulses ($\int U\,\mathrm{d}t$): It *sums* a series of impulses over a longer time of observation automatically.[C8.5] It is the limiting case of the ballistic galvanometer treated in Sects. 1.10 and 2.11. A mechanical analogy may be helpful:

In Fig. 8.11, a gravity pendulum is shown with its lower end dipping into a very viscous liquid, for example honey. This causes its motion to be strongly damped. We apply a hammer blow to the pendulum, giving it a mechanical impulse ($\int \boldsymbol{F}\,\mathrm{d}t$). The pendulum is deflected with a jerk and then practically stands still: due to its strong damping,

[C8.5] This is also the operating principle of the first telegraphic systems for use over long distances. C. F. GAUSS and W. WEBER set up a telegraphic connection in 1833 over a distance of ca. 1 km between the Göttingen Observatory and the Physical Academy. Positive and negative voltage impulses were used to cause a light pointer to deflect to the left or the right by a small amount. The apparatus can still be seen in the Historical Collection of the Physics Institute at the University of Göttingen.

[3] This is not yet the "creeping" pointer setting; see below!

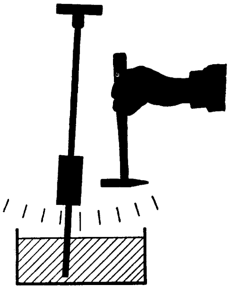


Figure 8.11 The operation of a creeping galvanometer

it can return to its rest position only very slowly, after several minutes. A second impulse (another hammer blow) would thus strike the pendulum at the end point of its first deflection, so the second deflection would add to the first. An impulse from the opposite direction (hammer blow from the left) would be correspondingly subtracted.

Creeping galvanometers were formerly used for measurements, mainly for registering voltage impulses. They can be calibrated as shown in Fig. 5.3, for example in volt second. As an example of the application of a creeping galvanometer, we investigate the effect of an iron core on the magnetic flux Φ of a current-carrying coil (see also Fig. 4.14). In Fig. 8.12, we can see that the coil is framed by an improvised induction loop. The pointer of the galvanometer is near the zero point of the scale (bottom image in Fig. 8.12). Now to the experiments:

1. The current in the field coil (about 3 A) is switched on. The galvanometer pointer begins to move to position a. This corresponds to 10^{-4} V s. That is the magnetic flux Φ of the empty field coil.

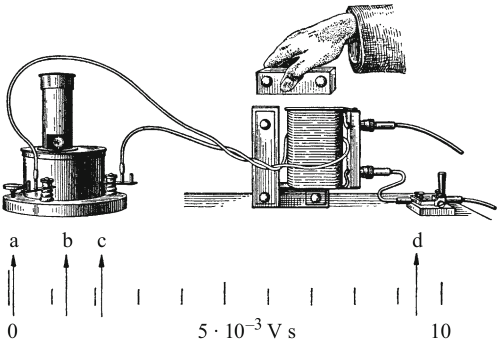


Figure 8.12 The effect of an "iron core" on the magnetic flux of a coil. The flux is measured using a creeping galvanometer, the same instrument as in Figs. 2.36, 2.38 etc., but now strongly damped by the small "external resistance" of the induction loop. The cross-sectional area of the iron core is around 25 cm^2. To amplify the deflections of the galvanometer, one could use several turns instead of only a single induction loop.

2. We put the field coil over one leg of the U-shaped iron core. The pointer moves to position b, since the flux Φ has increased to $1.3 \cdot 10^{-3}$ V s.

3. We bring an iron closure bar slowly into position on the U-shaped iron core and finally lay it firmly onto the core, closing the magnetic circuit. The pointer moves stepwise to position d, and the flux Φ has now reached a value of $9.4 \cdot 10^{-3}$ V s.

4. We switch off the current to the field coil, and the pointer of the galvanometer moves back to position c on the scale; the "remanent" magnetization of the iron core has a magnetic flux of $2.2 \cdot 10^{-3}$ V s. Finally, we remove the closure bar and the iron core. The pointer returns to the zero point. The field coil with no current and no iron core is once again free of magnetic flux.

A qualitative evaluation can be made readily using the simple model of molecular currents (Sect. 4.4). The magnetic field of the field coil aligns the magnetic fields of the molecular currents in the iron so that they are all parallel to each other and to the external field. Then the invisible ampere turns within the iron add to the visible ones in the field coil, greatly increasing the strength of the magnetic flux density $\boldsymbol{B}$. We will treat this topic quantitatively later, in Chap. 14 (Sect. 14.11, ferromagnetism). For the next chapters, the experience gained from these measurements will suffice: *The magnetic flux Φ of a current-carrying coil can be increased nearly 100-fold by inserting an iron core. Furthermore, it can be conveniently varied by changing the magnetic circuit of the core.*

8.5 The Magnetic Dipole Moment $\boldsymbol{m}$

The simplest and most convenient indicator of a magnetic field is no doubt a compass needle. The magnetic field exerts a torque $\boldsymbol{M}_{\text{mech}}$ on a suitably-mounted bar magnet. The latter can also be replaced by a current-carrying coil, for example as in Fig. 1.10. How does this torque come about, how can we describe it quantitatively? We give the answer first for the case of a current-carrying coil. The planes of its windings are parallel to the direction of the magnetic field, as shown in Fig. 8.13. We show only one turn of the windings instead of the whole coil; for simplicity, it has a rectangular cross section. Of the four sides of the coil, two (the vertical sides) are perpendicular and the two others are parallel to the field. Therefore, a force $F = BIl$ acts on the first two. The two forces act as a force couple on the lever arms r and produce the torque:[C8.6]

$$M_{\text{mech}} = BIl2r = BIA \tag{8.9}$$

A is the area of the windings, independently of the shape of the coil (rectangular, circular etc.).

C8.6. To avoid confusion with the magnetization M (Chap. 14), the symbol for the mechanical torque M is denoted by a subscript "mech", as already used in Sect. 3.9. For many of the equations in the following paragraphs, just the magnitudes are sufficient. The directions can be read off from the figures.

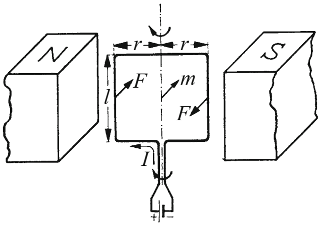


Figure 8.13 The origin of the magnetic moment (the current I is flowing in the conventional direction)

Now we introduce a new quantity, the *magnetic moment*. We define it by the equation

$$\boldsymbol{m} = I\boldsymbol{A} \tag{8.10}$$

$$\left(\text{unit: } 1\,\text{A}\,\text{m}^2\right).$$

The magnetic moment $\boldsymbol{m}$ is a vector quantity. Its direction is perpendicular to the area enclosed by the current path, i.e. parallel to the surface vector $\boldsymbol{A}$ which denotes the orientation and magnitude of the current path. An observer looking in the direction of $\boldsymbol{A}$ sees the current I flowing in a clockwise direction (or: the curved fingers of the right hand point in the direction of current flow, when the thumb is parallel to $\boldsymbol{A}$ and $\boldsymbol{m}$). Then the torque is given by

$$\boldsymbol{M}_{\text{mech}} = \boldsymbol{m} \times \boldsymbol{B}\,. \tag{8.11}$$

$$\left(\text{For example: } M_{\text{mech}} \text{ in N m, } m \text{ in A m}^2,\ B \text{ in V s/m}^2\right).$$

The vector product describes the torque for every orientation of $\boldsymbol{m}$ relative to $\boldsymbol{B}$. In Eq. (8.11), the vector $\boldsymbol{B}$ corresponds to the vector $\boldsymbol{E}$ in the analogous equation (3.25) for an electric dipole in an electric field.

Usually, instead of a *single* rectangular winding, one has coils with many turns of arbitrary shapes (long or flat, with a constant cross-sectional area A as in cylindrical coils (solenoids), or varying as in multilayered coils, especially in flat coils). For these cases, we recall for the second time an experiment from mechanics. In Fig. 3.17, we saw a bar (vector $\boldsymbol{S}$) attached to the end of a spoke R. The bar was subject to a torque $\boldsymbol{S} \times \boldsymbol{F}$ produced by a pair of forces $\boldsymbol{F}$, $-\boldsymbol{F}$ (a force couple). The length of the spoke R plays no role here.

Analogously, we can simply *add* the torques that act on each of the individual windings of a coil, independently of their distances from the common axis; this is just the same as for an electric dipole moment, treated in Sect. 3.9. For the overall torque, we obtain

$$\boldsymbol{M}_{\text{mech}} = I\left(\sum \boldsymbol{A}_{\text{i}}\right) \times \boldsymbol{B}\,. \tag{8.12}$$

In the case of well-defined cylindrical coils with only a few layers in their windings, all N turns have practically the same area A and the

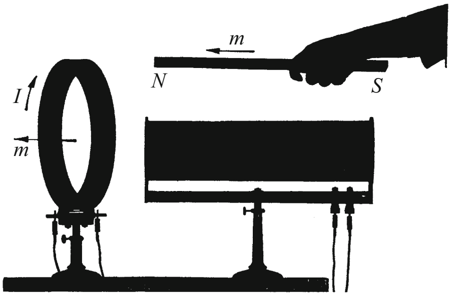


Figure 8.14 A bar magnet and two iron-free coils, all with the same magnetic moment, of magnitude $m \approx 34\,\mathrm{A\,m^2}$. The long coil has a diameter of 10.6 cm and 4300 turns; the flat coil is 25.4 cm in diameter and has 730 turns. The current in both coils is ≈ 0.9 A. The straight arrows show the directions of the three equal magnetic moments $\boldsymbol{m}$ and also the direction of the magnetic flux density $\boldsymbol{B}$ at the center of the flat coil. Looking in the direction of $\boldsymbol{m}$, we would see the current I circling clockwise (curved arrow, right-hand rule!). The north pole N of a compass needle would point to the geographic north pole of the earth.

same orientation. Therefore, their combined magnetic moment is

$$\boldsymbol{m} = IN\boldsymbol{A}\,. \tag{8.13}$$

Two examples of the magnetic moments of coils are shown in Fig. 8.14.

Permanent magnets of all kinds, and magnetized pieces of iron or other magnetic materials, are no different (outside their own volumes) from current-carrying coils or bundles of coils (Sect. 4.1); but the orbits of the circulating charges within these materials are invisible. As a result, we cannot compute the magnetic moment $\boldsymbol{m}$ of a permanent bar magnet or a similar object in the same way as that of a current-carrying coil (Eq. (8.12)). However, it can be measured using Eq. (8.11), in the simplest case with $\boldsymbol{m}$ perpendicular to $\boldsymbol{B}$,

$$m = \frac{M_{\mathrm{mech}}}{B}\,.$$

For this measurement, we mount the permanent magnet on bearings (like a compass needle) with minimal friction, in a horizontal plane. In equilibrium, that is $\boldsymbol{M}_{\mathrm{mech}} = 0$, its magnetic moment $\boldsymbol{m}$ orients itself parallel to $\boldsymbol{B}$ (see Eq. (8.11)). Then we use a calibrated torque (a spring balance F at the end of a lever arm r as in Fig. 8.15) to rotate the axis between the poles of the magnet until it is perpendicular to a homogeneous magnetic field of known flux density $\boldsymbol{B}$ (e.g. from a field coil). Figure 8.15 shows a measurement of this type using a bar magnet in the magnetic field of the earth.

Small torques $\boldsymbol{M}_{\mathrm{mech}}$ cannot be measured with great precision as the product of force times force arm; it is better to compute them from the oscillation period T of torsional oscillations. From Vol. 1, Eq. (6.13), we find the

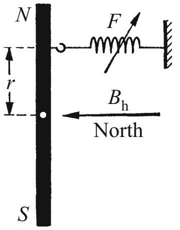


Figure 8.15 Measurement of the magnetic moment of a bar magnet, mounted horizontally and free to rotate in the earth's magnetic field. (A counter-torque $M_{\text{mech}} = rF$, for example $F = 7.8 \cdot 10^{-3}$ N with the lever arm $r = 0.1$ m, is produced by the spring, used as a force meter). The torque vector $\boldsymbol{M}_{\text{mech}}$ from the spring points perpendicular to the plane of the page and into it. The magnetic flux density of the horizontal component of the earth's field is $B_h = 2 \cdot 10^{-5}$ V s/m^2. Then we find $m = M_{\text{mech}}/B_h = 39$ A m^2.

ratio of the torque to the rotation angle, termed the *torsion coefficient*, to be

$$\frac{M_{\text{mech}}}{\alpha} = 4\pi^2 \frac{\Theta}{T^2} \tag{8.14}$$

(Θ = moment of inertia). If it is rotated out of its rest position through the *small* angle α, a freely suspended, horizontal bar magnet (compass) is subject to a restoring torque according to Eq. (8.11) of

$$M_{\text{mech}} = mB \sin\alpha \approx mB\alpha \,. \tag{8.15}$$

Equations (8.14) and (8.15) combined yield

$$m = \frac{4\pi^2 \Theta}{T^2 B} \,. \tag{8.16}$$

(For example T in s, Θ in kg m^2, for a bar magnet = (1/12) mass · (length)^2 (Vol. 1, Eq. (6.11)), B in V s/m^2).

An arbitrary magnetic object (current-carrying coil, compass needle, a paramagnetic molecule etc.) has a *magnetic moment* $\boldsymbol{m}$. We bring it into a magnetic field $\boldsymbol{B}$. Then the object will orient itself, presuming that it is free to move, so that its magnetic moment $\boldsymbol{m}$ is parallel to $\boldsymbol{B}$.

Applying an external torque that is opposed to the torque produced by $\boldsymbol{B}$, we can rotate $\boldsymbol{m}$ by an angle α relative to $\boldsymbol{B}$. This requires that the torque perform work:

$$W = mB \int_0^{\alpha} \sin\alpha \, \mathrm{d}\alpha = mB(1 - \cos\alpha) \,. \tag{8.17}$$

It is stored in the form of potential energy. For $\alpha = 180°$, i.e. an antiparallel orientation of $\boldsymbol{m}$ and $\boldsymbol{B}$, the work has its maximum value, $2mB$.

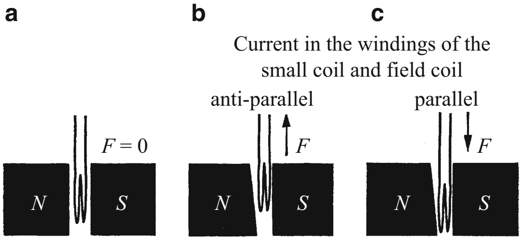


Figure 8.16 *Part a*: In a homogeneous field, a current-carrying coil, that is an object with a magnetic moment ***m***, is not acted on by a force. *Parts b and c*: In an inhomogeneous field, in contrast, forces act. This is at the same time a model of a diamagnetic substance (*Part b*) or a paramagnetic substance (*Part c*).

In an inhomogeneous magnetic field, in addition to the torque $\boldsymbol{M}_{\text{mech}}$, there is also a force $\boldsymbol{F}$. It pulls or pushes the object in the direction of the field gradient, e.g. $\partial \boldsymbol{B}/\partial x$. This important difference between homogeneous and inhomogeneous fields is elaborated in Fig. 8.16.

We will elucidate the origin and the magnitude of this force by referring to Fig. 8.17. We imagine the magnetic field to be perpendicular to the plane of the page and pointing towards the reader. The points where the field lines pass through the page are marked with dots. The field strength increases on going from top to bottom in the figure.

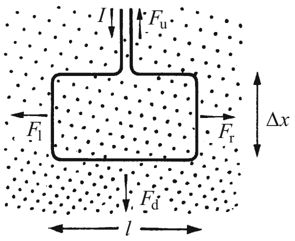
Our 'object with a magnetic moment ***m***' is a rectangular wire loop carrying a current I (its area is $A = l\Delta x$). Its magnetic moment $\boldsymbol{m}$ therefore points in the same direction as $\boldsymbol{B}$. The forces directed to the left and to the right, $\boldsymbol{F}_\text{l}$ and $\boldsymbol{F}_\text{r}$, cancel each other. The forces pulling upwards and downwards, $\boldsymbol{F}_\text{u}$ and $\boldsymbol{F}_\text{d}$, however, have different magnitudes. These are given by Eq. (8.1):

$$F_\text{u} = IlB \quad \text{and} \quad F_\text{d} = Il\left(B + \frac{\partial B}{\partial x}\Delta x\right) .$$

Therefore, the net force $F = F_\text{d} - F_\text{u}$ is pulling downwards; i.e.

$$F = Il\frac{\partial B}{\partial x}\Delta x = IA\frac{\partial B}{\partial x} ,$$

Figure 8.17 The derivation of Eq. (8.18). The current I flows in the conventional direction and the vector field ***B*** points perpendicular to the plane of the page towards the reader. When ***m*** and ***B*** are parallel, the resultant net force pulls downwards; when they are anti-parallel, it pulls upwards.



or, with Eq. (8.10),

$$F = m\frac{\partial B}{\partial x}\,. \tag{8.18}$$

This force pulls the object with the magnetic moment $\boldsymbol{m}$ into regions with a stronger or weaker field strength[4]. The sign is found as shown in Fig. 8.17. We can make use of Eq. (8.18) e.g. to determine an unknown field gradient $\partial B/\partial x$ using a test coil with a known magnetic moment $\boldsymbol{m}$.

A *numerical example*: In Fig. 8.16b and c, we had $m = 0.116\,\mathrm{A\cdot m^2}$ (i.e. 2 turns with an area of $20\,\mathrm{cm^2}$ carrying a current of 29 A); $F \approx 0.2\,\mathrm{N}$. Therefore, $\partial B/\partial x = 1.72\,\mathrm{V\,s/m^3}$.

8.6 The Localization of Magnetic Flux

In Fig. 4.7, the pole regions of a long current-carrying coil and a permanent magnet of similar shape made of a ceramic oxide material were compared. For both, the magnetic flux Φ could be localized using a wire induction loop (measured e.g. as in Fig. 5.8; the result is shown at the top of Fig. 8.18)[5]. For these two magnets, the *polar regions* can be schematically defined as in Fig. 8.20. When the magnetic flux is so strongly localized, we can apply a formal analogy between *magnetic flux* Φ and *electric charge* Q.

Suppose that in Fig. 8.13, the rectangular current path shown is one of the N turns of a solenoid whose axis is perpendicular to the plane of the page. It has a length l and is located in a homogeneous magnetic field of flux density $\boldsymbol{B}$. Its cross-sectional area is A and it is carrying the current I. This current produces the magnetic field strength $H' = NI/l$ inside the solenoid. In the homogeneous field $\boldsymbol{B}$, it is acted on by a torque $M_{\mathrm{mech}} = NIAB$ according to Eq. (8.12). This torque corresponds to a force couple, with a force $F = (NI/l)AB$ acting at each end of the coil as shown schematically in Fig. 8.19. With $NI/l = H' = B'/\mu_0$ and $B'A = \Phi'$, it then follows that (for simplicity we leave off the prime on Φ'):

$$\boldsymbol{F} = \Phi\boldsymbol{H}\,. \tag{8.19}$$

Here, $\boldsymbol{F}$ and $\boldsymbol{H}$ are vectors and Φ is a scalar quantity. Equation (8.19) corresponds formally to the equation

$$\boldsymbol{F} = Q\boldsymbol{E} \tag{3.5}$$

[4] Equation (8.18) also holds when the field gradient is parallel to the direction of the field. Suppose that in Fig. 3.18, $\boldsymbol{E}$ is replaced by $\boldsymbol{H}$ and the charge Q by the magnetic flux Φ, as explained in the following section.

[5] For bar magnets made of steel, which were formerly used extensively and can still be found today, the flux distribution is shown by the lower image in Fig. 8.18. There, one can localize the poles N and S at the "centers" of the shaded areas. In the case of magnetic fields from flat current-carrying coils (as in the left-hand image in Fig. 8.14), one can no longer speak of poles at all.

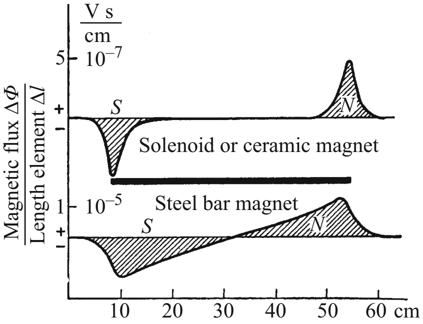


Figure 8.18 The distribution $\Delta\Phi/\Delta l$ of the magnetic flux Φ. In the upper part of the figure, a long current-carrying coil (solenoid) or a long bar magnet made of a magnetic ceramic oxide. Lower part: a bar magnet made of steel. An induction loop (as in Fig. 5.8) is slid stepwise along the length elements Δl and measures their contributions $\Delta\Phi$ to the magnetic flux Φ. (**Exercise 8.3**)

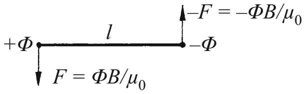


Figure 8.19 Schematic of a magnetic dipole in a homogeneous magnetic field. The field ***H*** or ***B*** lies in the plane of the page and points downwards, perpendicular to the line l.

for a charge Q in an *electric* field $\boldsymbol{E}$. By analogy, the magnetic flux Φ was previously termed the "magnetic charge". It was presumed to correspond in the magnetic case to a quantity of electricity or charge Q in an electric field $\boldsymbol{E}$.

The implementation of Eq. (3.5) with an electric field was treated in detail in Sect. 3.4. What was said there can be applied correspondingly to the implementation of Eq. (8.19) with a magnetic field, i.e. in particular: *For **H**, in* Eq. (8.19), *one must use the undisturbed magnitude, as measured before Φ was present.*

From Eq. (8.19), it follows that the magnitude of the torque which acts in Fig. 8.19 is given by

$$M_{\text{mech}} = Fl = \Phi H l = \frac{1}{\mu_0}\Phi B l$$

and that of the magnetic moment is

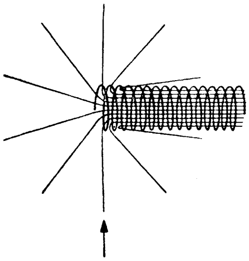
$$m = \frac{M_{\text{mech}}}{B} = \frac{1}{\mu_0}\Phi l\,. \tag{8.20}$$

We list some additional formulations which follow from the analogy between the magnetic flux and the electric charge:

1. *The magnetic field at a large distance from a magnetic pole which has the magnetic flux Φ.* Figure 8.20 shows a schematic drawing of

C8.7. See also the calculated field distribution in Comment C4.4.

Figure 8.20 The left end of a long, thin solenoid whose field lines emerge with approximately radial symmetry$^{C8.7}$



the field lines of a long solenoid (Fig. 4.4). For simplicity, we have drawn only its left-hand end in this figure.

At some distance from the polar region, the pattern of the field lines is to a good approximation radially symmetric (Fig. 8.20). The longer the bar magnet or coil, the more precise is this approximation. The magnetic flux distributes itself at large distances r symmetrically over a spherical surface $4\pi r^2$. Therefore, at a sufficiently large distance, we find for the magnitudes of $\boldsymbol{B}$ and $\boldsymbol{H}^{C8.8}$

C8.8. The behavior described by Eq. (8.21) can be experimentally determined with quantitative precision by using an induction coil (probe coil).

$$B_\mathrm{r} = \frac{\Phi}{4\pi r^2} \qquad \text{or} \qquad H_\mathrm{r} = \frac{\Phi}{4\pi\mu_0 r^2}\,, \tag{8.21}$$

once again completely analogous to the electric field of a point charge.

2. *The magnetic field directly in front of the flat face of a polar region.* In Fig. 5.8, we show the measurement of the magnetic flux Φ of a long coil. The measurement loop was near the center of the coil before it was pulled out; in Fig. 8.20, this step corresponds to the loop starting from far off to the right. It has cut through all the field lines in the process of being withdrawn from the coil.

In contrast to this step, we now place the measurement loop directly in front of the end of the coil, above the arrow. When it is pulled away, only the field lines to the left of the arrow pass through it, i.e. half of the total. That yields for the magnetic flux through the face of a coil end $\Phi_\mathrm{s} = \Phi/2$ (cf. also Fig. 8.18). Division by the cross-sectional area A of the coil gives the magnitudes of the fields B_s and H_s directly at the face of the coil; we find the values:$^{C8.9}$

C8.9. A different derivation of Eq. (8.22): In the interior of a long field coil, we find the magnetic flux Φ. If we now split the coil in half (in a thought experiment), then each of the newly-formed ends will make the same contribution to Φ, from symmetry. Note that only the axial component of $\boldsymbol{B}$ plays a role here.

$$B_\mathrm{s} = \frac{1}{2}\frac{\Phi}{A} \qquad \text{and} \qquad H_\mathrm{s} = \frac{1}{2\mu_0}\frac{\Phi}{A}\,. \tag{8.22}$$

3. *The magnetic field at a large distance R from an object with a magnetic moment $\boldsymbol{m}$.* Current-carrying coils (with or without an iron core) and permanent magnets can have the same magnetic moments $\boldsymbol{m}$ even if their shapes and compositions are quite different; we saw this in Fig. 8.14.

In the neighborhood of such coils or permanent magnets, the pattern of the field lines certainly depends on the shape of the magnetic

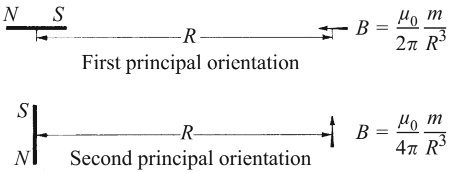

N S
R
$B = \frac{\mu_0}{2\pi}\frac{m}{R^3}$
First principal orientation
S
N
R
$B = \frac{\mu_0}{4\pi}\frac{m}{R^3}$
Second principal orientation

Figure 8.21 The flux density ***B*** at a large distance *R* from the center of a bar magnet or a coil with the magnetic moment ***m***

object. But, *at a sufficiently large distance, the fields* ***B*** *and* ***H*** *are determined only by the magnetic moment* ***m*** *of the object*. This is shown for the two *principal orientations* in Fig. 8.21. Here, the carrier of the magnetic moment is a small bar magnet with its north and south poles marked as *N* and *S*, often called a *magnetic dipole*.

Derivation: Each end of the bar (or coil) produces a flux density $B_\text{r} = \dfrac{\Phi}{4\pi R^2}$ at the point of observation according to Eq. (8.21). Only the difference of the two values is important, so that in the first principal orientation

$$B = \frac{\Phi}{4\pi}\left(\frac{1}{(R-l/2)^2} - \frac{1}{(R+l/2)^2}\right). \qquad (8.23)$$

When the distance R is sufficiently large compared to the length l of the bar or coil, we can neglect l^2 relative to R^2, and for the magnitude of B, we then obtain

$$B = \frac{1}{2\pi}\frac{\Phi l}{R^3} = \frac{\mu_0}{2\pi}\frac{m}{R^3}. \qquad (8.24)$$

Correspondingly, for the second principal orientation, we find

$$B = \frac{\mu_0}{4\pi}\frac{m}{R^3}. \qquad (8.25)$$

4. *The measurement of unknown magnetic moments using one of the principal orientations*. Equations (8.24) and (8.25) are important for measurements, in particular for the experimental determination of unknown magnetic moments ***m***. For this purpose, one measures ***B*** in one of the principal orientations, either directly with a probe coil (Sect. 5.4) or my making some sort of comparison with the known horizontal component of the flux density of the earth's field (e.g. $B_\text{h} = 0.2 \cdot 10^{-4}\,\text{V s/m}^2$ in Göttingen). For example, one orients the directions of B and B_h perpendicular to each other and finds the angle α between the directions of B_h and the vector sum of the two fields (Fig. 8.22) with the help of a compass needle. Then the field sought is given by $B = B_\text{h}\tan\alpha$. Using this value of B, one then computes the moment m from Eq. (8.25).C8.10

Compensation methods are often favored. In these, one allows a second, known magnetic moment to act on the compass needle in addition to the unknown moment (Fig. 8.23). The known moment is produced by a current-carrying coil of well-known dimensions. The magnetic moment of this "compensation coil" is computed using Eq. (8.13).

C8.10. In order to determine all the magnetic quantities by means of mechanical measurements, one would first have to measure the product mB_h (see Eq. (8.16)) by means of the experiment described in Fig. 8.15. Then one would find, as shown here, the angle α which the compass needle makes with B_h (independently of the magnetic moment of the compass needle). From Eq. (8.25), we then obtain

$$\tan\alpha = \frac{\mu_0 m}{4\pi R^3 B_\text{h}}.$$

Combining this with Eq. (8.16), it follows that

$$B_\text{h}^2 = \frac{\pi\Theta\mu_0}{T^2R^3\tan\alpha}.$$

This permits us to measure B_h without knowing the magnetic moment of either the bar magnet (coil) or of the compass needle; and from it, the other magnetic quantities follow (C. F. Gauss and W. Weber, 1832; see Wilfred Dudley Parkinson (1983), "Introduction to Geomagnetism" (Elsevier), p. 353 ff).

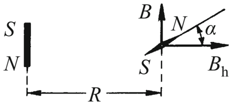


Figure 8.22 The determination of the flux density B of a dipole field in the second principal orientation, by making use of the known flux density of the horizontal component of the earth's field

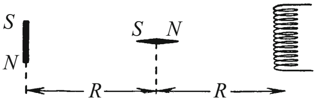


Figure 8.23 Measurement of an unknown magnetic moment by comparison with a coil of known magnetic moment m (null method). This is a schematic drawing (as is also Fig. 8.22); in reality, the distances R must be large compared to the dimensions of the carriers of the magnetic moments (the bar magnet S-N and the field coil).

5. *Forces between the planar, parallel faces of two neighboring magnetic poles*. One pole by itself produces the flux density

$$B_s = \frac{1}{2}\frac{\Phi}{A} \tag{8.22}$$

directly in front of its pole face. This field acts on the magnetic flux Φ of the other pole according to Eq. (8.19) with the force

$$F = \frac{1}{2\mu_0}\frac{\Phi^2}{A} = \frac{1}{2\mu_0}B^2 A\,. \tag{8.26}$$

This equation can be verified quite impressively by using a small electromagnet (a "pot-type electromagnet") with a diameter of only 5.5 cm (Fig. 8.24). When connected to a flashlight battery, it can pick up more than 100 kg[C8.11]

6. *The energy content of a homogeneous magnetic field of volume V*. In Fig. 8.25, the two faces of the magnetic poles approach each other over a distance Δx and can then lift a heavy load. In this process, a magnetic field of volume $V = A\Delta x$ is excluded. At the same time, the mechanical work[C8.12]

$$W = F\Delta x = \frac{1}{2\mu_0}B^2 A\Delta x = \frac{1}{2\mu_0}B^2 V\,. \tag{8.27}$$

was performed. Therefore, a homogeneous magnetic field of flux density B and volume V contains the energy

$$W_{\text{magn}} = \frac{1}{2\mu_0}B^2 V\,. \tag{8.28}$$

C8.11. To estimate the magnetic field H in the solenoid, one can use the expression for the field in the interior of a long coil or solenoid: $H = NI/l$. With $N = 500$, $I = 0.1$ A and $l = 1$ cm, we obtain $H = 5000$ A/m. From this value, with $\mu_0 = 4\pi \cdot 10^{-7}$ V s/A m, we find for the flux density B a value of about $6 \cdot 10^{-3}$ V s/m^2, more than two orders of magnitude smaller than the value measured in the induction experiment. This is a clear indication that H and B are not simply related by the factor μ_0 in the presence of magnetic material (here: the iron core). Also, B depends strongly on the geometry, i.e. here on the width of the gap **(Video 8.4)**. More details are given in Chap. 14.

C8.12. In this thought experiment, we have to take care that the flux density B is held constant in the space between the pole faces (see Comment C8.11). This can be fulfilled approximately by two permanent magnets at a sufficiently small spacing, so that stray fields can be neglected. A still more convincing demonstration of magnetic field energy will follow in Chap. 9.

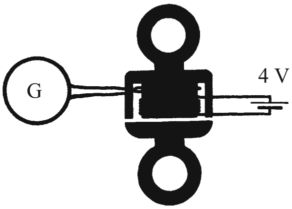
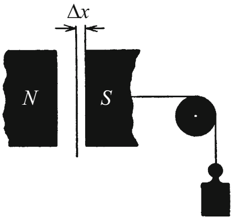


Figure 8.24 An electromagnet ("pot-type magnet") with a closed iron core. The field coil is at the center, and above it, an induction loop for measuring the flux density B. (The cross-sectional area A of the iron is $10\,\text{cm}^2 = 10^{-3}\,\text{m}^2$, $B = 2\,\text{V s/m}^2$, F calculated from Eq. (8.26) is $1.6 \cdot 10^3$ N). Using a flashlight battery as current source, the windings of the field coil should have around 500 turns. **(Video 8.4)**

Video 8.4: "Electromagnet"
http://tiny.cc/xbggoy
The strong attractive force of the electromagnet is demonstrated. It amounts to 530 N. In addition, it is shown how sensitively this force depends upon the width of the gap. A few sheets of paper are shoved between the two halves of the magnet. At a gap width of 0.4 mm, the force is only 15 N ! (see Sect. 14.6).

Figure 8.25 The calculation of the magnetic field energy



A *numerical example*: The highest flux densities B which can be obtained in iron cores are around $2.5\,\text{V s/m}^2$. Then in the field region between the poles, a magnetic field energy of ca. $2.5\,\text{W s/cm}^3$ is stored.[C8.13]

C8.13. Using superconducting magnet coils with flux densities of 5–$15\,\text{V s/m}^2$, energy densities of 10–90 W s/cm^3 can be obtained (for comparison: the stored energy density in Ni/Cd storage batteries is in the range of 180–300 W s/cm^3; see also the example in Vol. 1, Sect. 19.8, Point 1). Such superconducting magnetic energy storage devices (SMES) have been under development for some years. See e.g. P.J. Hall and E.J. Bain (2008), "Energy-storage technologies and electricity generation", *Energy Policy* **36** (12), pp. 4302–4309.

Exercises

8.1 For the experiment described in Comment C8.1, find the field B in which electrons that have been accelerated by a voltage of $U = 100\,\text{V}$ will follow a circular orbit of radius $r = 10\,\text{cm}$. (Sect. 8.1)

8.2

A flat coil carrying a current consists of 10 turns of wire with a cross-sectional area $A = 100\,\text{cm}^2$; it is in a magnetic field, as shown in Fig. 8.13. The current through the coil is $I = 10$ A and the flux density of the magnetic field is $B = 0.1$ T. How large is the torque M_{mech} which acts on the flat coil? (Sect. 8.5)

8.3 Estimate the magnetic flux Φ in the long coil used for the measurement results shown in Fig. 8.18, and compare it with the measurements given in Video 5.1 (Fig. 5.8), which were carried out on a different long coil. (Sect. 8.6)

8.4

Current is flowing through a helical conductor which is hanging loosely and vertically. Due to the current, it contracts along its long axis. Which force F would be necessary to keep it from contracting? The number density of its windings is $N/l = 2\,\text{cm}^{-1}$ when $I = 0$, the diameter of the helix is $2r = 5\,\text{cm}$, and the current strength is $I = 14\,\text{A}$. Start by determining the magnetic field energy W_{magn} in the helix. (Sect. 8.6)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_8) contains supplementary material, which is available to authorized users.

Applications of Induction, in Particular to Generators and Motors

9

9.1 Preliminary Note. General Remarks on Current Sources

Figure 9.1 serves to illustrate the general definition of the term *current source* or *generator*. A pair of condenser plates or "electrodes" A and C are connected through an ammeter. Between these electrodes are charges of both signs; we can think of them as localized on charge carriers. Two of them, a charge-carrier pair, are sketched in Fig. 9.1. The distance between the positive and the negative charges, measured along a horizontal line between the electrodes, can be increased by some sort of *charge-separation forces*. During their motion (not only when the charges reach the electrodes!), the ammeter indicates the flow of a current. In order to move the charges, the charge-separation forces must perform work. It is taken from a reservoir of mechanical, thermal or chemical energy.

If the outer circuit between C and A is interrupted, no more charges can flow between the plates. Then the charge-separation forces can for a time continue to transfer more charges to the two electrodes and thereby increase the voltage between C and A; but soon a limiting voltage, often called the *load-independent voltage*, will be reached, and it cannot be exceeded. The electric field between the plates itself produces forces on the charges between C and A and comes into equilibrium with the *charge-separation forces*, preventing further charging of the electrodes[1].

A modern sewing machine can be characterized by two innovations: the needle's eye at the tip of the needle, and the use of two independent threads. In a similar manner, the essential attributes of electrical machines can be summarized in a few sentences. The underlying physical principles and the decisive innovations are always

[1] All of these charge-separation forces were called "electromotive forces" in the past. However, that term was also used for the *voltages* that they produce, i.e. for the load-independent voltage of the current source, and it was devalued by this double usage. Besides that, it is much too long, so it is usually abbreviated as 'E.M.F.' In any case, one has to distinguish clearly between the charge-separation *forces* and *electrical quantities* such as the *voltage* which results from the charge separation.

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_9

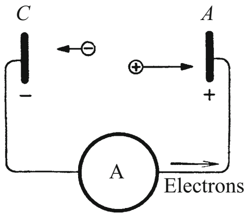


Figure 9.1 The definition of the term "current source". For demonstration experiments, we use two 'charge spoons' (cf. Fig. 2.11) as charge carriers (they are charged by an influence machine).

simple. The great achievements of modern electrical engineering are not a matter of physics, but rather of technology. Physical descriptions of these technical applications should be limited to a brief overview. In this chapter, we discuss the application of induction and the LORENTZ force to electric generators and motors.

9.2 Inductive Current Sources, Generators

We begin with the most important current sources in use today, *generators* which operate by *induction*. The *charge-separation forces* which they employ are produced by induction processes. We defined the terms 'charge-separation forces' and 'current source' on the basis of Fig. 9.1. We repeat that picture here in Fig. 9.2 with two additions: We adopt a viewpoint from within the 'black box' and suppose that it contains a magnetic field, perpendicular to the plane of the page, and also that the two electrodes C and A are connected by a conducting block (shaded). We can now separate the charges within this conductor in two ways and give them velocities which propel them towards the two electrodes:

1. We could *move the conductor* upwards in the direction of the arrow with a velocity $\boldsymbol{u}$, using it as a 'slider'. This will cause a *charge-separation force*, the LORENTZ force

$$\boldsymbol{F} = Q(\boldsymbol{u} \times \boldsymbol{B}) \qquad (7.5)$$

to act on the charges Q and move them in opposite directions.

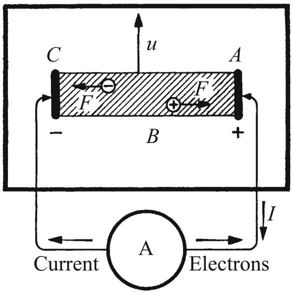
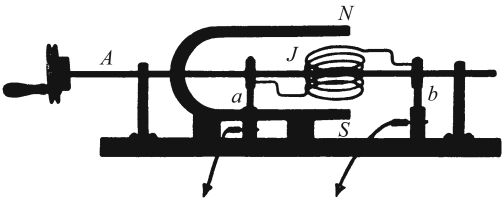


Figure 9.2 The definition of an "inductive" current source (the magnetic field points out of the page towards the reader and is perpendicular to the plane of the page; the current is flowing in the conventional direction, + to –)



To the voltmeter

Figure 9.3 An alternating-current generator with external magnetic poles

2. We could *change the magnetic flux density* $\boldsymbol{B}$ of the magnetic field which acts within the box. That would lead to an electric field around the closed current loop in Fig. 9.2 (cf. Fig. 6.2 and Fig. 5.5), and would move the charges between C and A by means of the forces $\boldsymbol{F}_+ = Q_+\boldsymbol{E}$ and $\boldsymbol{F}_- = Q_-\boldsymbol{E}$ towards the electrodes.

As a rule, both of these processes are applied simultaneously in order to use this device as a generator (current source) which produces a voltage between C and A. We will explain this by taking as examples several types of generators:

a) The *alternating-current generator with external magnetic poles* (Fig. 9.3). A coil J is rotated around an axle A in a magnetic field produced in some way. The ends of the coil are attached to two slip rings, which are electrically connected by two spring-loaded slip contacts or “brushes” a and b to the output terminals of the machine. The rotation of the coil represents the periodic repetition of a simple induction experiment. The induced voltage is “alternating”. Its time dependence can easily be registered using a voltmeter with a short response time (ca. 1 s) if the rotation of the coil is not too fast. In the special case of a homogeneous magnetic field and uniform rotation (Fig. 9.4a), this voltage curve is sinusoidal. Its frequency ν is equal to the rotational frequency.[C9.1]

For practical applications, the coil has an iron core (Fig. 9.5). The coil and the iron core together form the *rotor*. During their rotation, not only does the magnetic flux Φ through the rotor coil change, but

C9.1. A quantitative treatment is given in Sect. 5.6, in particular Eq. (5.10). More details on alternating current can be found in Sect. 10.3.

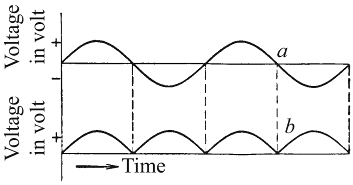


Figure 9.4 a): The sinusoidal voltage curve of an alternating-current or AC generator. b): The voltage curve of a direct-current generator with a simple coil rotor and a commutator (Fig. 9.6). The signs refer to the direction of the electric field between the output terminals.

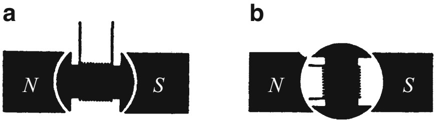


Figure 9.5 The iron cores of the stator *N*-*S* (field coils or permanent magnet) and of the rotor coils of a generator. At a, the magnetic flux Φ and the flux density $\boldsymbol{B}$ through the rotor are large, and at b, they are small.

Figure 9.6 A direct-current generator with a simple coil rotor *J*, commutator *C*, and a permanent-magnet stator

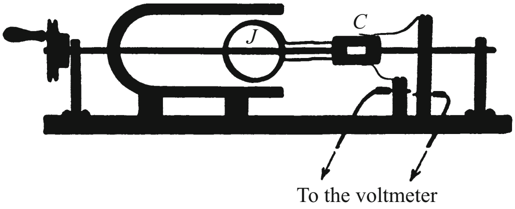
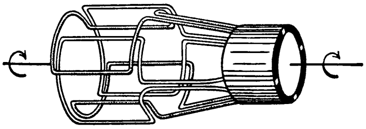


Figure 9.7 Cylindrical rotor with two pairs of coils and commutator

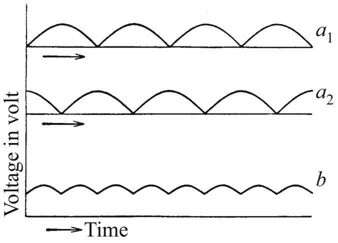


also the flux density $\boldsymbol{B}$, the latter due to the effective variation of the gap width (cf. Fig. 8.12).

b) The *direct-current generator*. Fig. 9.6 shows a demonstration model, again as a shadow projection. The slip rings of the alternating-current generator are now replaced by a simple switching device *C* (the “commutator”). It reverses the connections between the ends of the rotor coil and the output terminals of the generator after each half rotation. This folds the negative portions of the curve shown in Fig. 9.4a over to positive values; the voltage curve shown in Fig. 9.4b results. Its voltage varies between zero and a maximum value, but its sign always remains the same.

c) A *direct-current generator with a cylindrical rotor*. The arch-shaped voltage curve in Fig. 9.4b can be “smoothed”. Instead of a single coil *J*, one uses several, spaced at fixed angles relative to one another. This results in a *cylindrical rotor* rather than a *coil rotor*.

Figure 9.8 The voltage curve (*b*) of a cylindrical rotor with two pairs of coils, and how it is produced (superposition of a_1 and a_2)



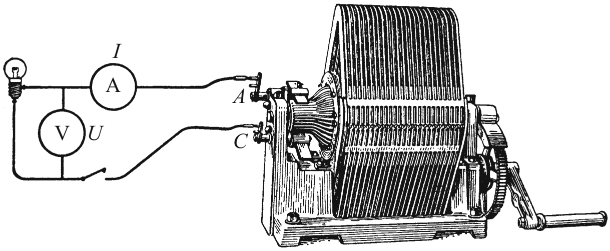
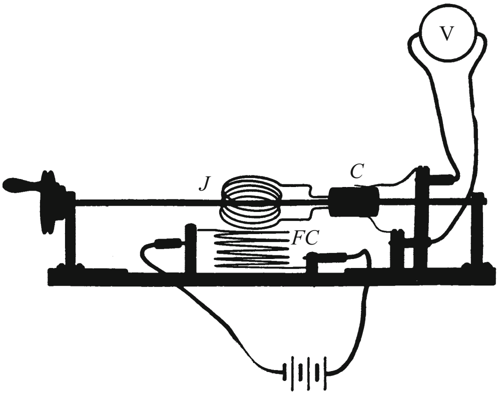


Figure 9.9 An old-fashioned direct-current generator with 2 × 25 permanent field magnets and a cylindrical rotor with 9 pairs of coils. At 8 A and 12 V, it can bring a 100-watt incandescent lamp to full brightness. One has to provide a muscle power of 8 A·12 V $\approx$ 100 W. The machine is "hard to crank". If the current is interrupted, however, it rotates freely, with hardly any resistance. This experiment demonstrates clearly that we should appreciate the energy content of a kilowatt hour and its commercial price ($\approx$ 30 Eurocent; see Comment C1.16.). As a demonstration object, the generator from an automobile engine is also suitable. However, as shown in the schematic in Fig. 9.10, it will require external excitation in the form of a current source for its field coils *FC*.

Figure 9.7 shows a schematic, with two pairs of coils and a fourfold commutator. In this example, two of the 'arch' curves from Fig. 9.4 (a_1, a_2) are superposed in an evident way. The result is the smoother direct current shown in curve 9.8b. Figure 9.9 shows a model of a direct-current generator with a cylindrical rotor which is suitable for use in lecture demonstrations.

d) The *direct-current dynamo*. The generators mentioned so far obtained their stator magnetic fields from permanent magnets. These permanent magnets can be replaced by current-carrying coils, so-called field coils (*FC* in Fig. 9.10). The current in the field coils may be supplied by some sort of auxiliary current source. Figure 9.10 shows a schematic of this external excitation. However, the genera-

Figure 9.10 A direct-current generator with external excitation



tor itself can also supply the current for its own field coils. This is the case with *dynamos*. Their principle is based on the presence of iron in the coils. When they begin to rotate, the weak remanent magnetic field of the iron (Fig. 14.7) induces an initial voltage in the rotor windings.

e) The *alternating-current generator with internal poles*. In the external-pole generator as described in a), the magnetic field which produces induction was fixed, and the rotor contained the induction coil J. In the case of the internal-pole generator, the opposite is true: The rotor consists of windings which carry a direct current. The fixed induction coil J is mounted on the stator. In practice, there are many coils which are arranged with radial symmetry. The rotor is often in the form of a flywheel, with the field coils around its circumference. The direct current for the field coils is provided by an auxiliary generator mounted on the same shaft as the main generator.

f) *Alternating-current generators with coil-free rotors*. In all the generators we have considered so far, the rotor, i.e. that part of the generator which rotates, carried coils. It is however possible to vary the magnetic flux within the induction coils J by means of a rotor without windings. Rotors of this type have the advantage that they are mechanically very stable and can therefore be operated at high rotational frequencies. Figure 9.11 shows a machine of this type. It is derived in a readily understandable manner from Fig. 8.12. The rotor consists in this model of a narrow, rectangular iron bar E, which causes the magnetic flux through the coils to vary, depending on its orientation.

For technical applications, one often replaces the permanent field magnets by electromagnets, i.e. coils carrying a direct current, with iron cores. Furthermore, all the parts are arranged with radial symmetry and repeat themselves many times around the circumference of the rotor and stator.

g) The *telephone as an alternating-current generator*. The essential point in the design of the alternating-current generator with coil-free rotors (Fig. 9.11) was the periodic variation of the iron-containing magnetic circuit. The rotation can be replaced by a back-and-forth oscillation (Fig. 9.12). M is an oscillating iron or steel diaphragm in place of the moving rotor. This again is only a technical variation on the experiment sketched in Fig. 9.11.

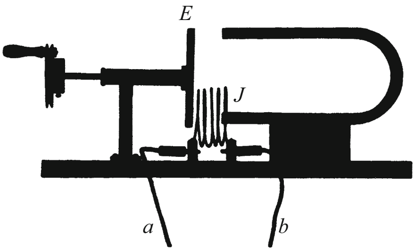


Figure 9.11 An alternating-current generator with a bar-magnet rotor E without windings. It is the 'closure bar' of the magnetic circuit formed by the horseshoe magnet.

Figure 9.12 Schematic of the telephone of Alexander GRAHAM BELL (1876)

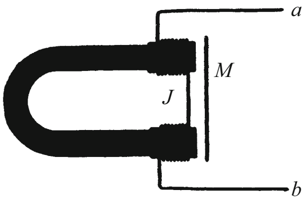
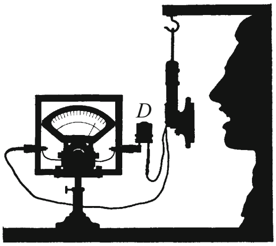


Figure 9.13 An antiquated telephone as an alternating-current generator (a rotating-coil ammeter is connected through a rectifier D)[C9.2]



C9.2. The telephone shown here is indeed quite antiquated, but even today, in the age of mobile telephones ("smartphones"), the same principle is still used in telephone receivers, loudspeakers and even the little sound converters in hearing aids. Microphones, in contrast, are increasingly based on capacitance or piezoelectric elements (see Sect. 3.10).

Figure 9.12 shows the schematic of a telephone transmitter. Here, it is interesting only as an alternating-current generator. Its function is to convert the mechanical energy of sound waves into electrical energy. To demonstrate this, we connect a telephone (Fig. 9.13) to an alternating-current ammmeter. When we sing into the diaphragm, we can observe weak currents of the order of 10^{-4} A. These alternating currents have the rhythm of the human voice. In earlier times, the audio currents were sent over long-distance telephone lines to the receiver telephone and converted there back to mechanical oscillations (sound waves). Figure 9.14 shows a sketch of a typical arrangement. Today, this setup is completely outmoded; human vocal cords are no longer used as a motor to drive an alternating-current generator. Instead, the pressure of the sound waves from the voice simply *controls* currents with the rhythm of speech (microphone)[2].

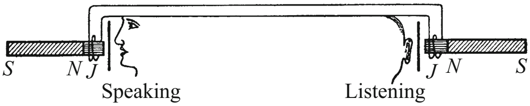


Figure 9.14 An old-fashioned connection of two telephones for long-distance calls (bar magnets instead of the horseshoe magnet as in Fig. 9.12)

[2] This kind of "control" was already utilized by the inventor of electric telephony, the teacher PHILIPP REIS (1861). The transmitter used by REIS was a microphone in today's terminology, with a vibrating contact made of platinum (instead of carbon particles as introduced by D. E. HUGHES in 1878). The telephone receiver used by REIS would today be termed a "magnetostriction receiver". In the first publication by REIS (1861), he closes with the words: "Until a practical application of the telephone becomes possible, there is still much work to be done. However, for physics, it is already interesting, in that it opens up a whole new field of research."

9.3 Electric Motors

All electric motors can in the end be reduced to the simple scheme shown in Fig. 9.15. We imagine that within the box outlined in black, there is a magnetic field with the flux density $\boldsymbol{B}$ perpendicular to the plane of the page, and the conductor C-A is brought into this field. By some means, we cause a current to flow through this conductor (for example from a current source operating at the voltage U_2). Then the conductor contains moving charges Q. Their velocities are indicated by the arrows u_+ and u_-.[C9.3] The magnetic field acts on these charges by exerting LORENTZ forces $\boldsymbol{F} = Q(\boldsymbol{u}\times\boldsymbol{B})$ (Eq. (7.5)). They cause the charges to move in the direction of the arrow a, carrying the conductor along with them (it is a simple 'slider'). In practice, a current-carrying coil is mounted as a "rotor" within the fixed magnetic field of the "stator". The forces acting on the rotor produce a *torque*.

C9.3. For details about the velocities in Fig. 9.15, see Sect. 8.2, in particular the footnote near the end of that section.

We give here two examples:

a) The *alternating-current synchronous motor*. This motor is in principle similar to an alternating-current (AC) generator. Figure 9.16 shows the same machine on the left as a generator and on the right as a motor. The rotor coils of the generator are turning at a frequency ν; thus it delivers alternating current with the same frequency ν. This current passes through the connecting leads 1, 2 into the rotor coils of the motor. There, it produces a torque which acts on the rotor coils. The direction of rotation depends on the direction of the current. Therefore, the torque must have the right direction at every position of the rotor to ensure that it continues to rotate. This can be guaranteed in a simple way:

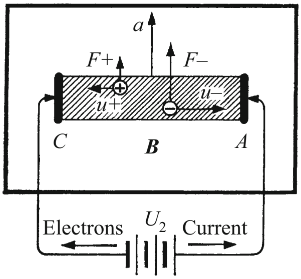


Figure 9.15 A schematic definition of an "electric motor" ($\boldsymbol{B}$ is perpendicular to the plane of the page, pointing outwards towards the reader)

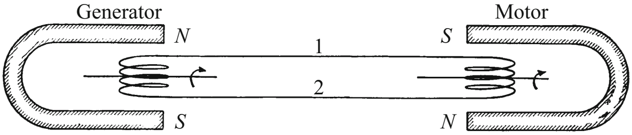


Figure 9.16 An alternating-current synchronous motor connected to an alternating-current generator with external poles

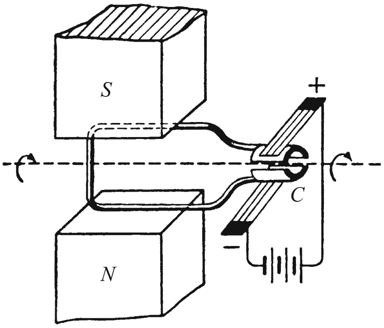


Figure 9.17 Schematic of a direct-current motor

In the rotor coils of the motor, the current produces a torque in the direction of the arrow shown in Fig. 9.16 at the moment represented there. After the time $T = 1/\nu$ (the rotational period), the current has again exactly the same direction and strength. If the rotor is then again at exactly the same position, then the torque will again act in the required direction. We have only to start up the rotor with the correct rotational frequency; thereafter, it will continue to rotate *synchronously* with the alternating current from the generator.

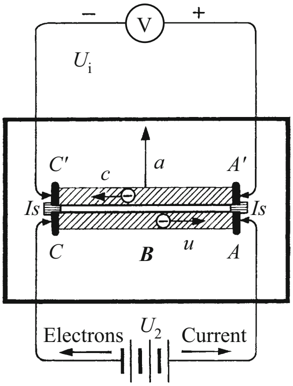
For a demonstration experiment, we wind a string around the axle of the motor in Fig. 9.16 and then pull it off, causing the rotor to begin turning like a child's spinning top. The alternating current from the generator has a frequency of $\nu = 50\,\text{Hz}$ and comes from the power grid (i.e. it is produced by a large AC generator and passed through the power transmission lines to our wall socket). In practice, there are several convenient ways of synchronizing the rotor's motion with the current when the motor is switched on. Alternating-current synchronous motors are widely used.

b) The *direct-current motor* is superficially similar to a DC generator. The simplified schematic of this motor is shown in Fig. 9.17. Its torque turns the rotor around its axle and pulls its coils until the plane of their windings is perpendicular to the page; then the direction of the current in the rotor coils is reversed, and so on after each half rotation. The reversals are accomplished automatically by the commutator *C*, which is rigidly attached to the rotor axle and acts as a switching device through its slip contacts or "brushes".

In this simple design, which is still used today, often in toys, the motor has a 'dead point'. It will not start running if the plane of its rotor coils is perpendicular to the magnetic field. In addition, its torque is not constant during a rotation. These problems are avoided by the cylindrical rotor. We already know its principle from DC generators (Fig. 9.7). Modern DC motors practically all use this design. The fields of the stator are always produced by current-carrying coils (electromagnets).

What determines the rotational frequency of the rotor? We repeat the schematic of a motor from Fig. 9.15 here in Fig. 9.18, but with two changes: First, for clarity, only the negative charges (electrons)

Figure 9.18 The induction process in the moving rotor of an electric motor. (*Is* are insulators, and the direction of ***B*** is perpendicular to the page, pointing towards the reader).



in the ‘rotor’ are shown. Second, we imagine that parallel to the current-carrying conductor C-A (which represents the rotor), there is a second conductor C'-A' of equal length. The two conductors are attached rigidly to each other, but are electrically insulated. The electrodes C' and A' are connected to a voltmeter.

When the current source U_2 is switched on, the conductor C-A (the “rotor of the motor”) begins to move in the direction of the arrow a (see Eq. (8.1)). This imparts a velocity in the direction of the arrow to the electrons in the parallel conductor C'-A'. As a result of this additional velocity in the magnetic field, a LORENTZ force acts on the electrons in the direction c (that is, opposite to their velocity $\boldsymbol{u}$ in the lower conductor!). This causes the voltmeter to register an induced voltage U_i (cf. Sect. 7.3).

Now we suppose that the conductors C'-A' and C-A are fused into *one single* conductor. Then we can see that the induced voltage U_i also occurs in the current-carrying conductor C-A. During its motion, the net voltage acting on the electrons in this conductor is $U_2 - U_i$. *In the limit that* $U_i = U_2$*, the current source can no longer deliver current to the ‘rotor’*. Then the acceleration due to the electromagnetic forces no longer occurs, and the conductor (‘motor rotor’) moves with a *constant* limiting speed in the direction of the arrow a. How could we increase this limiting speed? Either by increasing the voltage U_2 of the current source attached to the ‘rotor’, or by reducing the induced opposing voltage U_i, i.e. by *decreasing* the magnetic flux density ***B*** of the stator.

Both of these changes can be demonstrated on a motor with *external excitation* (Fig. 9.19), preferably a typical motor with a power rating of 1 kilowatt. When the current source is switched on with its voltage U_2, a strong ‘short-circuit current’ of many ampere flows through the rotor[3]. The resistance R_i of the coils of the rotor is not large, and the induced opposing voltage U_i which subtracts from the applied volt-

[3] In large electric motors, the windings of the coils and their power leads are in danger of overheating. This is prevented by using a “starter” (R_a in circuit b, Fig. 9.19). This variable resistor is gradually switched off during the run-up to

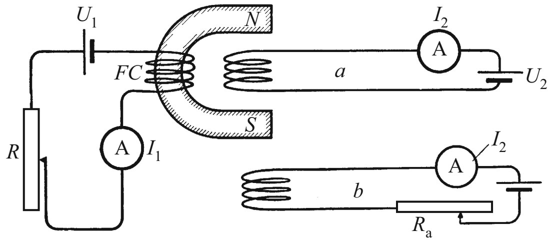


Figure 9.19 The induction process in the rotor of a direct-current motor with external excitation, known in electrical engineering as a "LEONARD circuit". The voltage U_2 of the current source can be varied in sign and magnitude in order to change the rotational frequency and direction of the motor; e.g. for driving a conveyor belt (the commutator is not shown here).

age U_2 is still very small. It grows only as the rotor gains speed; then the current through the rotor windings is controlled by the reduced voltage $U_2 - U_\text{i}$, and it approaches zero with increasing rotor speed. The limit $U_\text{i} = U_2$, when the rotor current drops to zero, can in practice never be reached; without current, the rotor can draw no more energy from the current source. It would have to continue rotating with only its stored kinetic energy. In reality, even when the rotor is not moving an external load, there are unavoidable frictional losses (and in addition JOULE heating in the windings). Therefore, even without a load, the rotor requires a certain energy input to maintain its rotational frequency. At least a small current must flow through its windings. Loading the motor, e.g. having it lift a weight, or braking the output shaft by manual friction, causes the current I_2 in the rotor windings to increase.

To conclude these experiments, we reduce the applied voltage U_2 to a very small value; it could be supplied for example by a 2-volt storage battery. Then the rotor reaches its constant, limiting rotational frequency at a very slow rotation rate. If we now *increase* its frequency manually, the ammeter will show a *reversal* of the direction of the current I_2. The voltage U_i induced in the rotor windings has then become *greater* than the voltage U_2 of the current source. The work performed by our hand is flowing as electrical energy back into the storage battery; the motor, now serving as a generator, is recharging its battery!

This experiment is very striking. It shows us that the technically so enormously important machines for converting electrical energy into mechanical energy are based physically solely on the LORENTZ force. In a generator, this force *accelerates* the electrons, producing an electric current and thereby converting mechanical work into electrical energy. In an electric motor, the same force converts the

speed of the motor, thus always keeping the current strength limited to acceptable values.

electrical energy into mechanical work. In addition, it brakes the electrons and thus limits the current in the rotor windings.

9.4 Three-Phase Motors for Alternating Current

A magnetic field whose direction is rotating, for short a 'rotating field', was already treated in some detail using Fig. 8.8. Such a rotating field can be produced be superposing two phase-shifted alternating currents. We make use of the general scheme shown in Vol. 1, Sect. 4.5. Here, we recall the essential points in Fig. 9.20.

Figure 9.21 shows a shadow projection of an AC generator on the left. Its rotor consists of two iron-core coils, J_1 and J_2. They are mutually displaced by 90°. The left-hand coil, which is just horizontal in the figure, appears to be a circular disk due to the foreshortened perspective. The ends of the two coils are connected to slip rings. The spring contacts ("brushes") a and b (and a' and b') collect the two alternating currents. These have a phase shift of 90° relative to one another. They are carried to the right in the figure, to two magnet coils that are mounted on an iron yoke perpendicular to each other and are split in the center. In their common center space, a rotating magnetic field is generated. To detect this field, an "induction rotor"

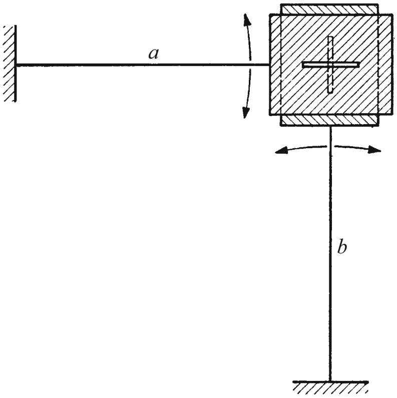


Figure 9.20 Production of *circular* mechanical vibrations by superposition of two perpendicular *linear* vibrations of the same frequency. Two long leaf springs a and b carry metal cards with slits parallel to the long axes of the springs. The overlapping opening of the two slits allows light to pass through. When projected, the spot of light follows a circular orbit when the springs are vibrating with the same amplitude and with a phase lag of one-quarter period ($\hat{=}$ 90°) after appropriate excitation. The diagonal of this circular orbit rotates like the spoke of a wheel. In a rotating-field motor, the direction of the magnetic field is represented by the diagonal. **(Video 9.1)**

Video 9.1:
"Circular Vibrations"
http://tiny.cc/qcggoy

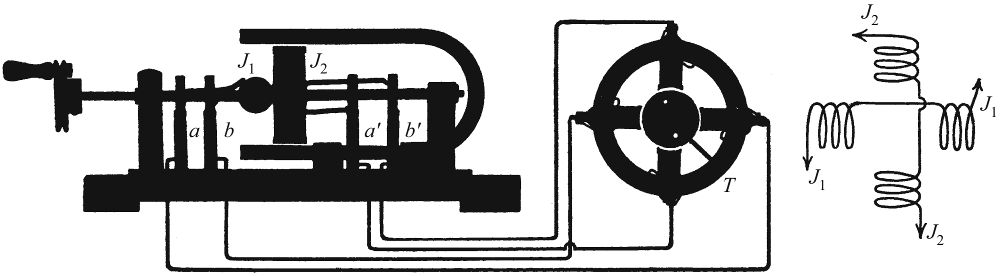


Figure 9.21 A demonstration model of a two-phase rotating-field generator and a rotating-field motor with an iron disk as rotor (compare Fig. 8.8)

as shown in Fig. 8.8, e.g. in the form of a metal disk, is mounted on an axle. Its axis of rotation is perpendicular to the plane of the page. In the figure, T is the mount for the bearings of this axle. The crossed magnet coils and the induction rotor together form a rotating-field motor.

Rotating-field motors are extremely important for practical applications. They can be constructed with a nearly ideal simplicity for power outputs of up to several kilowatt. They start up with a strong torque, without a starter resistor or capacitor (initially, they have a large slip (Sect. 8.3, Point 5)). Their rotational frequency is to a great extent independent of their loading. Apart from the slip, this rotational frequency is equal to the frequency of the alternating current or, with suitable modifications, to an integral fraction of its frequency.

We can distinguish between single-, two- and three-phase rotating-field motors. Figure 9.21 shows a two-phase motor. It requires four lead wires and is seldom used today.

A three-phase motor with so-called "three-phase current": Imagine that in Fig. 9.21, there were *three* rotor winding coils J, spaced at 120° intervals around the rotor axle. Correspondingly, in the right-hand part of Fig. 9.21, there would be *three* stator coils oriented at 120° intervals around the yoke. Then with three alternating currents, shifted by 120° on the time axis relative to each other, we would likewise obtain a rotating field or a circularly-polarized magnetic field. This would require six lead wires, but with a clever arrangement, two each can be connected in pairs so that only three wires are necessary. One can see these three wires in long-distance transmission lines.[C9.4]

The single-phase motor requires only two lead wires; it is fed with normal alternating current. The second, phase-shifted AC current which is required to generate a rotating field is produced within the motor itself by making use of some technical tricks. It must be phase shifted by 90° relative to the first AC current. The principle of this design will be explained later in Fig. 10.13.

C9.4. For a discussion of three-phase electrical power, see for example https://en.wikipedia.org/wiki/Three-phase_electric_power.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_9) contains supplementary material, which is available to authorized users.

10 The Inertia of the Magnetic Field. Alternating Current

10.1 Self-Inductance and the Inductance L

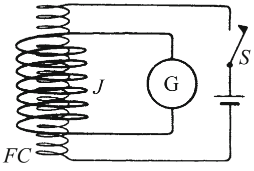
Self-inductance[1] refers to a special form of induction processes. Knowledge of this phenomenon is of great importance for a modern understanding of electromagnetism.

In demonstrating induction phenomena, we used among others the experimental setup sketched in Fig. 10.1. The current-carrying coil *FC* produces a magnetic field. A change in this field, caused for example by interrupting the current through the coil, induces a voltage impulse in the induction coil J, measured e.g. in volt second.

Now, the magnetic field penetrates not only the induction coil J, but also the field coil *FC* itself. *Therefore, any change in the field will also induce voltages in the field coil.* This is called self-inductance. In self-inductance, the time-varying magnetic field induces a voltage in the very coils which produce it.

Another derivation: Suppose that the field and induction coils in Fig. 10.1 are wound parallel on the same coil form, i.e. with the same dimensions. Then the two parallel wires could be fused together after winding in a 'thought experiment'. The two coils would become one! (Compare Fig. 9.18 and its explanation).

Figure 10.1 Schematic of an induction experiment



[1] Discovered by JOSEPH HENRY, 1832 (a watchmaker who later became professor of physics at Princeton University).

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_10

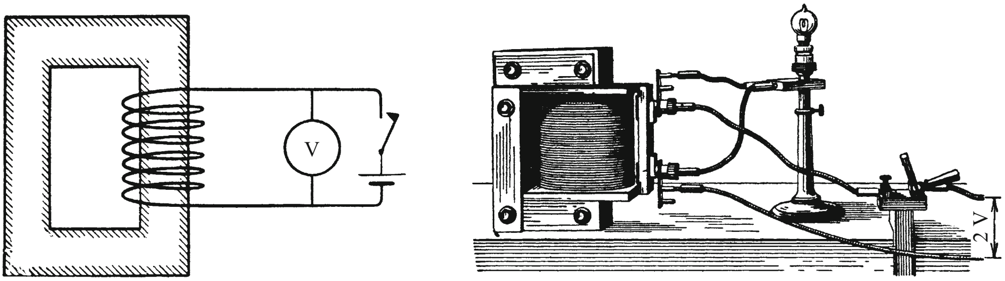


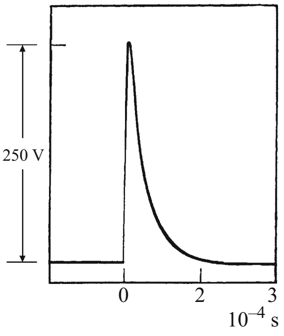
Figure 10.2 Demonstration of the voltage impulse resulting from self-inductance; at left, using a voltmeter; at the right, with a small light bulb (the inductance L of the coil is a few tenths of a henry (V s/A)). The time dependence of the voltage impulse can be displayed with an oscilloscope (Fig. 10.3).

To demonstrate self-inductance, we use the setup shown in Fig. 10.2: a wire coil with around 300 turns in its windings. To increase the magnitude of the voltage impulses, the coil is wound around a closed, rectangular iron core ('yoke').[C10.1] The ends of the coil are connected to a storage battery and, in parallel, a rotating-coil voltmeter. The voltmeter initially indicates the output voltage of 2 V from the storage battery. When we interrupt the current by opening the switch, the magnetic field decays rapidly. At the same time, the voltmeter shows a strong impulse deflection of up to ca. 20 V. The voltage thus attains briefly a much higher value than the original voltage that was applied to the coil, as a result of its self-inductance (compare Fig. 10.3). We could also replace the voltmeter by a 6-volt light bulb (Fig. 10.2, right). Its filament just glows weakly with a dull red light as long as only the battery voltage is present; but on opening the switch, we see it flash brightly to white heat: *Self-inductance releases energy, visible throughout the lecture hall, and it can only have been stored in the magnetic field of the coil.*

C10.1. See also Fig. 8.12. The effects of ferromagnetic materials on magnetic fields will be treated in more detail in Chap. 14.

According to the law of induction (e.g. Eq. (5.6)), the voltage impulse $\int U\,\mathrm{d}t$ induced in a coil depends on two factors: First, the change in the magnetic field, that is ΔH (or ΔB), and second, the dimensions and form of the coil. ΔH depends on ΔI, the change in the current

Figure 10.3 A voltage impulse due to self-inductance



strength during the process. We can thus write[C10.2]

$$\int U\,\mathrm{d}t = L\,\Delta I\,. \tag{10.1}$$

The proportionality constant L is called the *inductance*. We thus define the

$$\text{Inductance } L = \frac{\text{Induced voltage impulse}}{\text{Current change } \Delta I}\,. \tag{10.2}$$

The unit of this quantity is found to be 1 V s/A, termed 1 henry (H) or, equivalently, 1 Wb/A (weber per ampere).

The inductance is easy to calculate for a long, empty coil (solenoid) with a homogeneous magnetic field: We first consider the solenoid to be a field coil; it produces the field strength

$$H = \frac{NI}{l}\,. \tag{4.1}$$

The change ΔH in this field then causes the induced voltage impulse:

$$\int U\,\mathrm{d}t = \mu_0 N_\mathrm{J} A \frac{N\,\Delta I}{l}\,. \tag{5.1}$$

N_J is the number of turns in the windings of the inductance coil; here, it is the same as N, the number of turns in the field coil. Then we find

$$\int U\,\mathrm{d}t = \frac{\mu_0 N^2 A}{l}\,\Delta I\,. \tag{10.3}$$

Comparison to Eq. (10.1) yields the *inductance L* of the solenoid, which we were seeking:

$$L = \frac{\mu_0 N^2 A}{l}\,. \tag{10.4}$$

C10.2. As we did in Chap. 5 in our first discussion of induction, we here initially ignore the question of the sign in treating self-inductance. Thus, the equations in this section are all to be understood as giving magnitudes only. The signs will be dealt with in the following section.

10.2 The Inertia of the Magnetic Field as a Result of Self-Inductance

In demonstrating self-inductance, we have thus far ignored the sign of the induced voltage impulse. We now take it into account, which will lead us to a deeper understanding of the phenomenon of self-inductance.

We repeat the experiment as shown in Fig. 10.4. In the left-hand image, the voltmeter first indicates the output voltage of the battery (2 V) as a deflection to the left. The small fraction of the electrons which flows through the voltmeter is moving in the direction of the

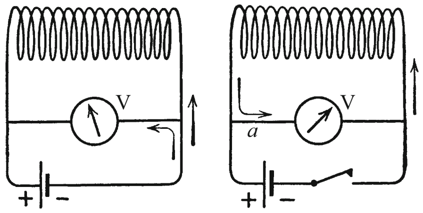


Figure 10.4 The inertia of the electric current in a coil (the arrows indicate the direction of flow of the electrons)

curved arrow. In the right-hand image, the battery has just been switched off. The large impulse deflection of the voltmeter goes to the right on its scale. The current through the voltmeter it thus now flowing in the opposite direction. Therefore, the current through the coil must also continue to flow for a time in the original direction even without an external current source, and this causes negative charges to collect at point a. The current and its resulting magnetic field exhibit *inertia*. They behave in an analogous manner to a massive body in motion, or a rotating flywheel.

We briefly recall an example of mechanical inertia: In Fig. 10.5, at left, we see a current of water moving around a closed circuit of piping, driven by the pump P. The Hg manometer between the points a and b indicates a pressure drop to the left, corresponding to the direction of flow and the resistance (friction) in the pipes. In the right-hand image, the pump has been switched out of the circuit by closing the valve H. Due to its inertia, the water continues to flow for a time in the direction of the arrow, so that the manometer now shows a strong deflection to the right. This principle was used to construct a technical application, the water lifting device known historically as the “hydraulic ram” (or “hydram”) (J. M. MONTGOLFIER, 1796).

A moving body or a flywheel show their inertia not only when their motion is slowed by braking, but also when they are initially set in motion. This process also requires a finite time. The same is true

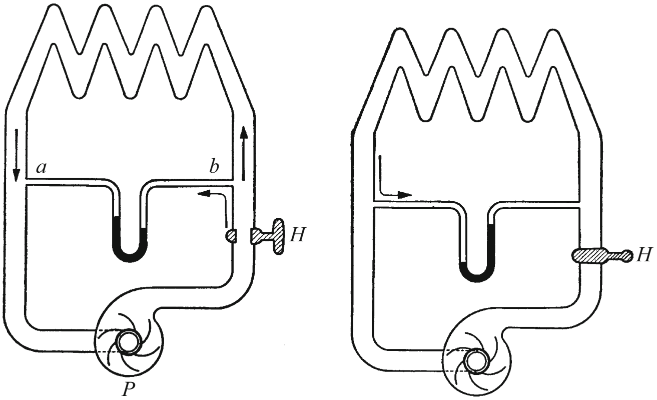


Figure 10.5 The inertia of a current of water in a pipe circuit

of a current and its magnetic field. This can be shown by an important and striking experiment (Fig. 10.6). The voltage U is again supplied by a storage battery (2 V). The ammeter A is a rotating-coil instrument with a fast response time (less than 1 s). The large coil, wound with thick wires, surrounds a closed iron yoke (cf. the scale drawing). When the switch 1 is closed, the pointer of the ammeter immediately starts to move; but it advances only slowly. After a minute, it is still creeping visibly upwards. Only after 1–1/2 minutes have the current and its magnetic field reached their full values; this demonstrates their considerable inertia.

After the maximum values of the current and the field have been attained, we short-circuit the battery with switch 2 and immediately take it out of the circuit by opening switch 1. We again can observe the inertia of the current and its magnetic field. Even after a minute, the ammeter is still showing a clear deflection from zero. These experiments are always rather surprising. We normally associate electrical processes in everyday life with instantaneous, momentary or timeless phenomena.

"These experiments are always rather surprising. We normally associate electrical processes in everyday life with instantaneous, momentary or timeless phenomena."

We have described this fundamental fact, the *inertia of currents and their magnetic fields*, intentionally here from a purely empirical point of view. Retrospectively, we can see that it is a simple result of LENZ's law: Let us take the second case as an example. There, we short-circuited the current source and then removed it from the circuit. In an ideal conductor, without electrical resistance, the current would simply continue to flow indefinitely. In fact, however, the best commercially-available wires have a finite resistance R, so that the current is consumed by quasi-frictional forces (JOULE heating, Sect. 1.12).C10.3

C10.3. An exception to this rule is exhibited by *superconductors*. In superconducting wires, induced currents can be maintained and observed for many years. The phenomenon of superconductivity is found (at sufficiently low temperatures) in most metals and a large number of compounds. It is characterized by the fact that below a certain temperature T_C, the electrical resistance of the material becomes zero. POHL described this in earlier editions. For a basic introduction to the subject, see W. Buckel and R. Kleiner, "*Superconductivity, Fundamentals and Applications*", 2nd edition, Wiley-VHC, Weinheim (2004).

This *gradual reduction* of the current is the cause of the induction process. The induced voltage must therefore *oppose* the decrease of the current, according to LENZ's law. A portion of the kinetic energy lost by the electrons through "friction" is replaced at the cost of the magnetic field energy, thus delaying the reduction of the current.

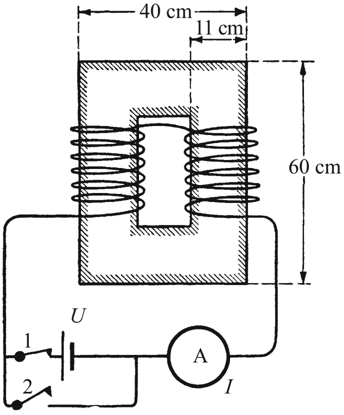
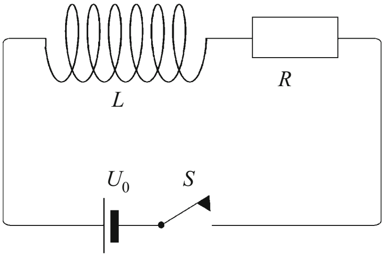


Figure 10.6 Switching on and off of a magnetic field in a coil requires some time **(Video 10.1)**

Video 10.1:
"The Inertia of a Magnetic Field"
http://tiny.cc/wcggoy
See Comment C10.5.

Figure 10.7 The derivation of Eq. (10.7)



For a quantitative description of the process, taking the sign found experimentally correctly into account, we rewrite Eq. (10.3):

$$\int U \,\mathrm{d}t = -L\,(I_2 - I_1)\,, \tag{10.5}$$

or, in differential form,

$$U = -L\,\frac{\mathrm{d}I}{\mathrm{d}t}\,. \tag{10.6}$$

The signs correspond to LENZ's law (Sect. 8.3); i.e. when the current is increasing ($I_2 > I_1$ or $\mathrm{d}I/\mathrm{d}t$ positive), the induced voltage opposes the current, while a decreasing current ($I_2 < I_1$ or $\mathrm{d}I/\mathrm{d}t$ negative) is accompanied by a voltage in the same direction as the current. See also Sect. 5.6.

For further considerations, we make use of the series circuit sketched in Fig. 10.7, with the usual trick that we draw the inductance L and the resistance R as if they were spatially separated (although in fact they are both properties of the same coil).

After the switch S is closed, the applied voltage U_0 drops across L and R; we thus have[C10.4]

$$U_0 = U_\mathrm{L} + U_\mathrm{R} = L\frac{\mathrm{d}I}{\mathrm{d}t} + RI\,. \tag{10.7}$$

C10.4. The voltage U_L compensates the voltage U induced in the coil:

$$U_\mathrm{L} = -U = L\,\frac{\mathrm{d}I}{\mathrm{d}t}$$

(see Sect. 10.4).

The solution of this differential equation is given by

$$I = \frac{U_0}{R}\left(1 - e^{-\frac{R}{L}t}\right)\,, \tag{10.8}$$

as one can readily convince oneself by substituting the solution (10.8) into Eq. (10.7). The time $\tau_\mathrm{r} = L/R$ is called the *relaxation time*. The current $U_0/R = I_\mathrm{max}$ is the saturation value reached only after several relaxation times.

If we now short-circuit the battery and remove it from the circuit, so that $U_0 = 0$, the solution of Eq. (10.7) becomes

$$I = I_\mathrm{max} e^{-\frac{R}{L}t}\,. \tag{10.9}$$

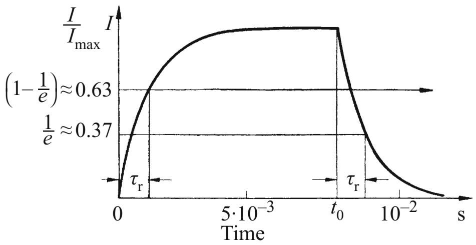


Figure 10.8 The time dependence of the current accompanying the increase and decrease of a magnetic field (Example: $L = 10^{-1}$ V s/A, $R = 10^2\,\Omega$, $\tau_r = 10^{-3}$ s (see also Fig. 2.51)).

Equations (10.8) and (10.9) are shown graphically in Fig. 10.8.[C10.5]

C10.5. For the derivation of Eqns. (10.8) and (10.9), we have assumed that the coil was empty (no iron core!). In the experiment described in Fig. 10.6 **(Video 10.1)**, this is however not the case. In that experiment, the coil was mounted on a closed iron yoke. As a result, the rise and subsequent fall of the current were delayed considerably. The strong effect of the iron core in the coil is especially clear at the end of the video, when after reversing the current, its rise is much slower than at the beginning of the experiment. The reason for this is the reversal of the magnetization of the iron (see Chap. 14, especially Fig. 14.7).

The energy W stored in the magnetic field can be calculated from the JOULE heating within the resistor R after switching off the battery (at the time t_0). We find (from Sect. 1.12):

$$\frac{\mathrm{d}W}{\mathrm{d}t} = I^2R = \left(\frac{U_0}{R}\right)^2 R\,e^{-\frac{2R}{L}t}\,. \qquad (10.10)$$

Integration gives

$$W = \frac{U_0^2}{R}\int_{t_0}^{\infty} e^{-\frac{2R}{L}t}\mathrm{d}t = \frac{1}{2}LI_{\text{max}}^2\,; \qquad (10.11)$$

that is, in a coil of inductance L carrying a current I, the stored magnetic energy is

$$W = \frac{1}{2}LI^2 \qquad (10.12)$$

(e.g. W in W s, L in V s/A, I in A).

Making use of Eqns. (10.4) and (4.1), we obtain from this the expression for the energy stored in a magnetic field of volume V that we had previously found in Chap. 8:

$$W_{\text{magn}} = \frac{\mu_0}{2}H^2V = \frac{1}{2\mu_0}B^2V\,. \qquad (8.28)$$

The inertia of the magnetic field plays a decisive role in all applications of electric currents where their strengths and directions vary.

Qualitatively, we can demonstrate two significant effects by making use of a periodically interrupted direct current (DC). In Fig. 10.9, the current from a 2-volt storage battery divides along two branches of the circuit, each with a light bulb. The left-hand branch contains in addition a coil with an iron core, while the right-hand branch has only

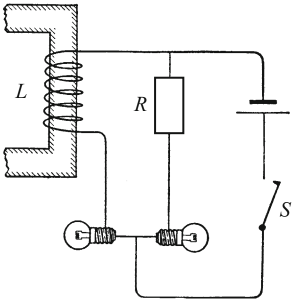


Figure 10.9 A demonstration of inductive resistance or *inductive reactance* using a periodically interrupted direct current

a short piece of wire with the same resistance as the coil (ca. 0.3 Ω). With a constant current strength, both branches are equivalent and both light bulbs glow with the same brightness. When the switch S is periodically opened and closed, the situation is quite different (the switch is briefly opened at time intervals of T, $1/T\; = $ frequency ν):

1. With a low frequency, both lamps still show the same brightness, but the left-hand lamp is delayed by about one second relative to the right-hand one. Its current lags behind the voltage. It requires nearly a second to build up its magnetic field.

2. As the frequency is increased, the time is no longer sufficient for the magnetic field to reach its full value. The left-hand lamp becomes dimmer and dimmer. At frequencies above 1 Hz, it stays quite dark. Thus, the coil has an *inductive resistance*, and it increases with increasing frequency.[C10.6]

C10.6. This is a model of a so-called *low-pass filter*, which allows only low frequencies to pass through.

In these experiments with periodically-interrupted direct current, all of the energy provided by the current source to build up the magnetic field is lost. In Fig. 10.9, when the switch is opened, the "sluggish" current in the coil flows through both lamps and converts the magnetic field energy there into heat. (Without the wire of resistance R and the right-hand lamp, this would occur instead in the form of an electric arc between the switch contacts (switch arcing)!). We now turn to sinusoidal ("alternating") currents.

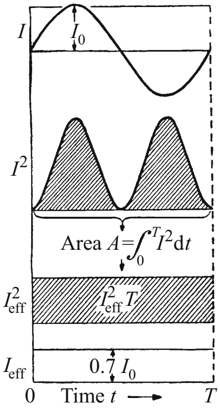
10.3 Alternating Current: A Quantitative Discussion

For a quantitative treatment of alternating current (AC), we choose the simplest form, i.e. a sinusoidal waveform. Alternating currents of more complex forms can always be constructed by superposing sinusoidal waveforms. The formalism of Fourier analysis, discussed in Vol. 1, Sect. 11.3, can be applied quite generally to the description of alternating currents.

For sinusoidal alternating currents and voltages, we have

$$I = I_0 \sin \omega t \quad \text{and} \quad U = U_0 \sin \omega t\,. \tag{10.13}$$

Figure 10.10 The definition of the effective current of a sinusoidal alternating current



Here, I and U are the momentary values (at time t) of the current and the voltage, I_0 and U_0 are their amplitudes, that is their maximum values, and $\omega = 2\pi\nu$ is their circular frequency (cf. Vol. 1, Sect. 4.3).

The momentary values of the current and the voltage can be observed and measured using instruments with a sufficiently short response time, for example oscilloscopes. In general, one does not measure or quote these momentary values, but rather the *effective* values, as their *time-averaged* values are called. The time-averaged value over a sine function would of course simply be equal to zero. Therefore, one calculates the average values of the *squares* of the time functions and defines their effective values by means of the equations

$$I_{\mathrm{eff}} = \left[\frac{1}{T}\int_0^T I^2\,\mathrm{d}t\right]^{1/2} \quad \text{and} \quad U_{\mathrm{eff}} = \left[\frac{1}{T}\int_0^T U^2\,\mathrm{d}t\right]^{1/2} \qquad (10.14)$$

$$(T = \text{period} = 1/\nu).$$

They are illustrated in Fig. 10.10. These effective values are also called "root-mean-square" (rms) values after their calculation procedure (Eq. (10.14)). For sinusoidal currents and voltages, one finds

$$I_{\mathrm{eff}} = \frac{I_0}{\sqrt{2}} \quad \text{and} \quad U_{\mathrm{eff}} = \frac{U_0}{\sqrt{2}}\,. \qquad (10.15)$$

The effective values of the current and the voltage are thus proportional to their amplitudes. This definition corresponds to the values of the direct current and voltage which would give the same average JOULE heating in an OHMic resistor.

10.4 Coils in Alternating-Current Circuits

The effects of self-inductance in a single cable can often be neglected. In such an "inductance-free cable", current and voltage would have the same phase. At every moment, the voltage as given by OHM's law (for short the "OHMic voltage" $U_R = IR$) is sufficient to maintain the current I along the conductors.

However, it is frequently not possible to neglect the self-inductance of the conductor in alternating-current circuits, especially when they contain coils. Then, to maintain the current, in addition to the OHMic voltage U_R, an additional *inductive* voltage U_L is required to compensate the induced voltage U_{ind}. From Eq. (10.6), we find

$$U_L = -U_{\text{ind}} = L\frac{dI}{dt}\,. \tag{10.16}$$

In this case, one draws the coil in a circuit diagram in two parts, for example as in Fig. 10.11: In the upper part of the circuit, only the OHMic resistance determines the voltage (middle curve), while in the lower part (bottom curve), only the self-inductance acts.

For a sinusoidal alternating current, Eq. (10.16) takes the form

$$U_L = L\omega I_0 \cos \omega t = L\omega I_0 \sin(\omega t + 90°)\,. \tag{10.17}$$

Then the amplitude of the inductive voltage is given by

$$U_{L,0} = I_0\omega L = I_0 2\pi\nu L \tag{10.18}$$

with the significant new result that the voltage amplitude $U_{L,0}$ leads the current amplitude I_0 by 90° on the time axis. The two voltage

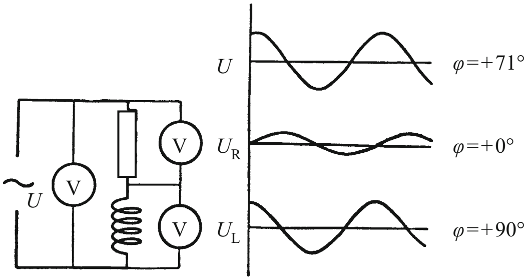


Figure 10.11 A series circuit consisting of a conductor with only OHMic resistance (above and middle curve) and a coil with only inductive resistance (below and bottom curve). The amplitude ratios and phase angles are similar to Fig. 10.12. The voltage U_R has the same phase as the current. (Numerical example: $\nu = 50\,\text{Hz}$, $L/R = 9.2 \cdot 10^{-3}\,\text{s}$). The top curve shows the overall voltage $U(t)$.

amplitudes required to maintain the current amplitude I_0, i.e. $U_{\mathrm{R},0}$ and $U_{\mathrm{L},0}$, must thus be combined to give a resultant voltage amplitude U_0.C10.7

C10.7. For this "combination" of the voltage amplitudes, the functions $U_\mathrm{R} = U_{\mathrm{R},0} \sin\omega t$ and $U_\mathrm{L} = U_{\mathrm{L},0} \cos\omega t$ (Eq. (10.17)) must be added. The result is again a harmonic function of the same period, but shifted by a phase angle φ, i.e. it leads the function U_R (Eq. (10.20)). The resulting amplitude can be found in Eq. (10.19). These results can be most clearly shown in a so-called phasor diagram (or vector diagram) as in Fig. 10.12. This has the advantage that other formulas for AC resistances can be most simply derived mathematically.

This can be represented graphically in a so-called *phasor diagram* or *vector diagram* as in Fig. 10.12. For the amplitude of the resultant voltage, we find

$$U_0 = I_0\sqrt{R^2 + (\omega L)^2}\,. \tag{10.19}$$

It leads the current amplitude I_0 in time by the phase angle φ. The phase angle is given by

$$\tan\varphi = \frac{\omega L}{R}\,. \tag{10.20}$$

The quotient

$$\frac{U_0}{I_0} = Z = \sqrt{R^2 + (\omega L)^2} \tag{10.21}$$

is called the overall AC resistance or *impedance* Z of the circuit. It is *not a constant* for alternating currents, but rather it *increases with the AC frequency* ν.

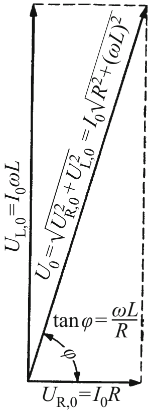
When the product ωL is large, the impedance Z can be orders of magnitude greater than the constant OHMic DC resistance R, as may be seen from Eq. (10.21). In such cases, R can be neglected in Eq. (10.21) relative to ωL. Then, only the *inductive* or AC resistance (also referred to as the *inductive reactance*) remains:

$$\frac{U_{\mathrm{L},0}}{I_0} = \omega L\,. \tag{10.22}$$

To summarize: The voltage

$$U = U_0 \sin\omega t \tag{10.23}$$

Figure 10.12 The calculation of the AC resistance or impedance from a "phasor diagram"C10.8


$U_{\mathrm{L},0}=I_0\omega L$
$U_0=\sqrt{U_{\mathrm{R},0}^2+U_{\mathrm{L},0}^2}=I_0\sqrt{R^2+(\omega L)^2}$
$\tan\varphi=\frac{\omega L}{R}$
φ
$U_{\mathrm{R},0}=I_0R$

C10.8. In a phasor diagram, the amplitudes of the sinusoidal voltages (or currents) are represented by the lengths of the lines (vectors). φ is the phase angle relative to a quantity which is the same in all parts of the AC circuit. In Fig. 10.12, this is the current, and φ is the angle defined by Eq. (10.24); here, it thus indicates by how much the voltage is leading the current. For calculations, the lines or 'phasors' in the diagrams can be treated like vectors. See also Comment C10.10.

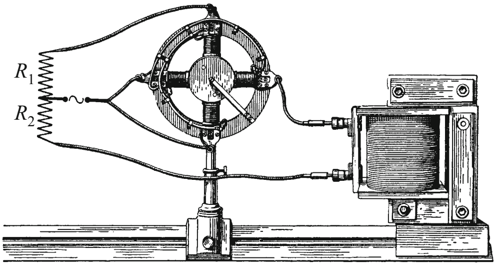


Figure 10.13 Demonstration of a phase shift by producing a rotating field (current ≈ 10^{-1} A. R_1 and R_2 are incandescent lamps used as ballast resistors instead of rheostats and ammeters, ∼ is an AC current source, $\nu = 50\,\text{Hz}$) **(Video 10.2)**

Video 10.2:
"Rotating Magnetic Field"
http://tiny.cc/8cggoy
First, a "null experiment" is shown: An OHMic resistor in place of the coil produces no motion of the iron disk. When the resistor is then exchanged for a coil or choke (inductance $L \approx 1\,\text{H}$), the disk begins to rotate to the right: The B field rotates in a clockwise sense. If a condenser (capacitance $C = 10\,\mu\text{F}$, see Sect. 10.5) is inserted instead of the coil, the disk rotates, like the magnetic field, in a counter-clockwise sense. The incandescent lamps used as ballast resistors are of limited usefulness as ammeters. The different impedances explain the different rotational frequencies that are observed in the two experiments.

produces a current

$$I = I_0 \sin(\omega t - \varphi) \tag{10.24}$$

with $I_0 = U_0/Z$. In the circuit shown in Fig. 10.11, the phase angle φ (Eq. (10.20) and Fig. 10.12) is positive.

The phase shift φ between the momentary values of alternating current and voltage is a favorite subject for fascinating demonstration experiments. A convenient setup for demonstrating the phase shift is shown in Fig. 10.13. The current from an AC current source (∼) is divided equally between two branches which each contain a pair of coils; these are mounted perpendicular to each other. One of these current branches also contains a coil with a large inductance. As a result of the phase shift, a *rotating magnetic field* is produced (Sect. 9.4, Fig. 9.21). A metal disk rotates in this field as rotor. (A practical application of this principle is found in the widely-used so-called "split-pole motors"[C10.9]) .

C10.9. The split poles of an AC magnet each carry a copper ring on one side ("shaded poles"). This delays changes in the magnetic flux. A rotating magnetic field results.

10.5 Condensers in Alternating-Current Circuits

In the experiment shown in Fig. 10.13, we exchange the coil for a condenser ($C \approx 10^{-5}\,\text{F}$). We again observe a *rotating magnetic field*, but its sense of rotation is opposite to that observed with the coil **(Video 10.2)**. We can draw two conclusions from this. First: *Alternating current* is not blocked by a condenser; it passes through as a *displacement current* (Sect. 6.4). Second: Between the current and the voltage, we again find a *phase shift* of 90°, but now, the current leads the voltage.

For a quantitative treatment, we assume a sinusoidal alternating voltage

$$U = U_0 \sin \omega t \tag{10.25}$$

($\omega = 2\pi\nu$ = circular frequency, ν = mechanical frequency).

Suppose that the condenser has a capacitance of C (and thus when a voltage U is applied to it, it carries a charge of $Q = CU$). Then at every time t, we find for the charging or discharging current

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t} = C\frac{\mathrm{d}U}{\mathrm{d}t}\,. \tag{10.26}$$

In this equation, $\mathrm{d}U/\mathrm{d}t = \omega U_0 \cos\omega t = \omega U_0 \sin(\omega t + 90°)$, and thus

$$I = C\omega U_0 \sin(\omega t + 90°)\,, \quad (\varphi = -90°, \quad \text{see Eq. (10.24)})\,. \tag{10.27}$$

Therefore, for the amplitude of the current which is passing through the condenser (the displacement current), we have

$$I_0 = \omega C U_0 = 2\pi\nu C U_0\,, \tag{10.28}$$

but with an important additional fact: The current leads the voltage by 90° (compare later to the schematic in Fig. 10.15). The quotient

$$\frac{U_0}{I_0} = \frac{1}{\omega C} \tag{10.29}$$

is the *capacitive* or *AC* resistance of the condenser, also called the *capacitive reactance*.

Numerical example: $\nu = 50\,\mathrm{Hz}$, $C = 10^{-5}\,\mathrm{F}$, $U_0/I_0 = 3.2 \cdot 10^4\,\Omega$.

10.6 Coils and Condensers in Series in Alternating-Current Circuits

For a conductor (resistor) which has only OHMic resistance, we have:

$$U_{\mathrm{R},0} = I_0 R\,, \tag{1.2}$$

and for a conductor with only inductive resistance,

$$U_{\mathrm{L},0} = I_0 \omega L\,, \tag{10.18}$$

while for a condenser through which a displacement current I_0 is flowing, we find

$$U_{\mathrm{C},0} = \frac{I_0}{\omega C} \tag{10.29}$$

(the index 0 again denotes the amplitudes).

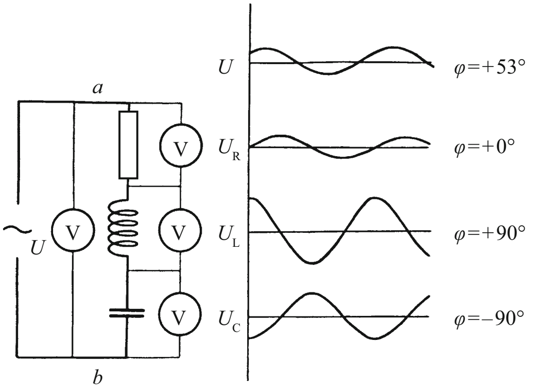


Figure 10.14 A series circuit with a condenser, a coil and an OHMic resistor. (Numerical example: $\nu = 50\,\text{Hz}$, $L/R = 1.44 \cdot 10^{-2}\,\text{s}$, $1/RC = 10^3\,\text{Hz}$. Note that the voltage U_R has the same phase as the current in this series circuit. In the example shown, the current lags by 53° behind the applied voltage $U(t)$.)

When these three components are connected in series (Fig. 10.14), the three voltages just mentioned add with their respective phases to give the total voltage U_0:

$$U_0 = I_0 \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2} \tag{10.30}$$

(Fig. 10.15). Relative to the current I_0, it is phase shifted by the phase angle φ; and relative to the voltage U_C across the condenser, it is phase shifted by $(\varphi + 90°)$. We can find φ according to Fig. 10.15, right side:

$$\tan\varphi = \frac{\omega L - \frac{1}{\omega C}}{R} \tag{10.31}$$

(see Eq. (10.24) for the definition of the phase angle φ. A graph of $(\varphi + 90°)$ is given later in Fig. 11.18c).

For each pair of values of L and C, there is a special frequency ν_0 at which the inductive resistance ωL and the capacitive resistance $1/(\omega C)$ become equal. When we set these two quantities equal, we obtain the *resonance frequency*

$$\nu_0 = \frac{1}{2\pi\sqrt{LC}}\,. \tag{10.32}$$

Experimentally, this "resonance of a series circuit" is demonstrated in Fig. 10.16. The OHMic resistor R and the inductive resistance ωL are combined in a single coil (with an iron core). The partial *voltages* $(U_R + U_L)$ and U_C have larger amplitudes and effective values than the total (applied) voltage U_0. For this reason, one often speaks

Figure 10.15 The calculation of the AC resistance of the series circuit in Fig. 10.14

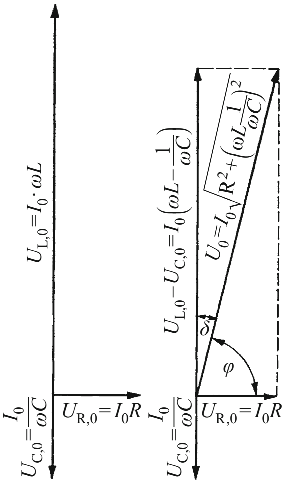

$U_{L,0}=I_0\cdot\omega L$
$U_{C,0}=\frac{I_0}{\omega C}$
$U_{R,0}=I_0R$
$U_{C,0}=\frac{I_0}{\omega C}$
$U_{R,0}=I_0R$
$U_{L,0}-U_{C,0}=I_0\left(\omega L-\frac{1}{\omega C}\right)$
$U_0=I_0\sqrt{R^2+\left(\omega L-\frac{1}{\omega C}\right)^2}$
δ
φ

Figure 10.16 An example of *voltage resonance* in a series circuit (ν_0 = 500 Hz; the current source is an AC power supply with variable frequency (signal generator); the coil has a closed iron yoke, and $L \approx 37$ H, $R = 1.1 \cdot 10^4\ \Omega$. A rotary variable condenser is used to adjust the resonance frequency, with $C_{\max} \approx 3 \cdot 10^{-9}$ F. Static voltmeters (electrometers E) are used (Sect. 1.6)).

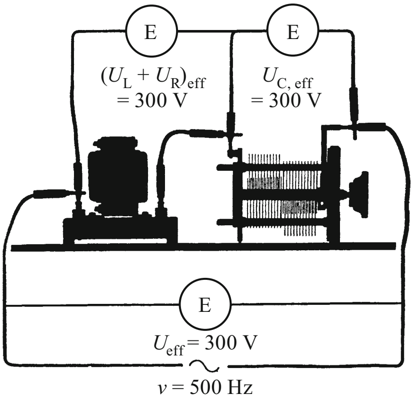
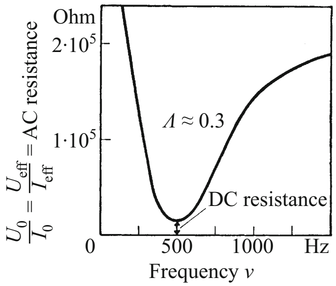

E
E
$(U_L + U_R)_{\text{eff}}$ = 300 V
$U_{C,\text{eff}}$ = 300 V
E
U_{eff} = 300 V
ν = 500 Hz

Figure 10.17 The AC resistance of the series circuit as a function of the AC frequency (Eq. (10.30)). (The experimental data are those in Fig. 10.16, and Λ is the logarithmic decrement (Vol. 1, Sect. 11.10); see also Sect. 11.7).



of a *voltage resonance*. At resonance in a series circuit, the overall resistance U_0/I_0 has its minimum value (Fig. 10.17).

In every electrical circuit, due to its OHMic resistances R, electrical energy is converted into heat. The power consumed in this way is $\dot{W} = I^2R$. Other losses can also occur, especially in iron-containing coils due to eddy currents and magnetization reversals (Sect. 14.4). All the losses, i.e. the total power lost, is attributed to an effective resistance R', which is larger than the DC resistance. We thus define $R' = (\sum \dot{W})/I^2$. In the ideal limiting case $R' = 0$, the two voltages $(U_\mathrm{L} + U_\mathrm{R})$ and U_C would be phase shifted relative to each other by exactly 180°, and at resonance, both would increase without limit (resonance catastrophe!).

10.7 Coils and Condensers in Parallel in Alternating-Current Circuits

The resistor, the coil and the condenser can also be connected together in a parallel circuit (Fig. 10.18). Then we obtain quite different results for the amplitude U_0 from that of the series circuit (Eqns. (10.30) through (10.32)), by adding together the partial currents. We find[C10.10]

C10.10. The mathematical derivation of Eqns. (10.33) and (10.34) can again be carried out with the aid of a phasor diagram, this time however for the amplitudes of the *currents*. An equivalent method is the complex-number formalism. A compact introduction is given for example at http://www.physics.byu.edu/faculty/peatross/homework/complex145.pdf .

$$U_0 = I_0 \frac{\sqrt{R^2 + (\omega L)^2}}{\omega C \sqrt{R^2 + (\omega L - \frac{1}{\omega C})^2}}\,, \tag{10.33}$$

and for the phase angle φ between U and I (its sign is as defined in Eq. (10.24)),

$$\tan\varphi = \frac{\omega L}{R}(1 - \omega^2 LC) - \omega RC\,. \tag{10.34}$$

In the limit of very high frequencies ($\omega \to \infty$), we have $\varphi = -90°$. At resonance, that is for $\varphi = 0$, we have

$$\omega = \omega_0 \sqrt{1 - \frac{R^2 C}{L}}\,, \quad \omega_0 = \frac{1}{\sqrt{LC}}\,. \tag{10.35}$$

This equation is the same as Eq. (10.32) only in the special case that $R^2 C/L \ll 1$ (which is however often fulfilled). Resonance can be demonstrated experimentally with the circuit shown in Fig. 10.19. Here, as in the series circuit, the Ohmic resistance R and the inductive resistance ωL are again located in a single coil. The two partial *currents*, i.e. the current I_L flowing through the coil and the displacement current I_C passing through the condenser can have much greater amplitudes and effective values than the overall current I_0. Therefore, one often speaks here of a *current resonance*.

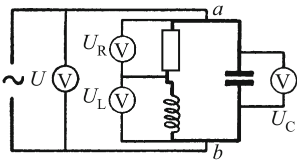


Figure 10.18 A coil and an Ohmic resistor in series are connected in parallel with a condenser. The voltages are measured as effective values, or else as momentary values with an oscilloscope.

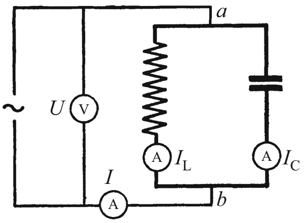


Figure 10.19 An example of a *current resonance* in a *parallel circuit* ($\nu = 50\,\text{Hz}$; a technical paper condenser was used, with $C = 3.7 \cdot 10^{-6}\,\text{F}$; $R = 38\,\Omega$, $L = 2.7\,\text{H}$. (The zig-zag symbol used here refers to a coil which also has an OHMic resistance.) (**Exercises 10.10, 10.12**).

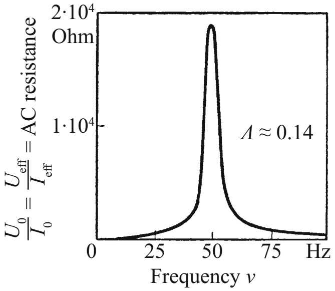


Figure 10.20 The AC resistance of a parallel circuit as a function of the AC frequency (Eq. (10.33)). (The experimental data are from Fig. 10.19; Λ is the logarithmic decrement (Vol. 1, Sect. 11.10); see also Sect. 11.7) (**Exercise 10.11**).

As a result of the losses mentioned in the fine print in Sect. 10.6, the current amplitude I_0 may become very small, but never zero. The phase difference between I_L and I_C can approach the value 180° arbitrarily closely, but never reaches it.

At resonance, the overall resistance $U_0/I_0 = U_\text{eff}/I_\text{eff}$ of the parallel circuit has its maximum value (Fig. 10.20) (leading to the name 'rejection circuit').

10.8 Power in Alternating-Current Circuits

For the power $\dot{W}$ of any electric current, we have found $\dot{W} = IU$. In the case of alternating current, both I and U are periodic functions of the time. Furthermore, in general there is a phase difference φ between them. In the simplest case, i.e. with sinusoidal alternating currents and voltages, Eqns. (10.23) and (10.24) apply, that is, the power is given by

$$\dot{W} = IU = I_0 U_0 \sin \omega t \sin(\omega t - \varphi)\,, \qquad (10.36)$$

or, after an elementary rearrangement,

$$\dot{W} = \frac{1}{2} I_0 U_0 [\cos\varphi - \cos(2\omega t - \varphi)]\,. \qquad (10.37)$$

In words: The power $\dot{W}$ of an AC circuit consists of two parts: the first, $\frac{1}{2} I_0 U_0 \cos\varphi = I_\text{eff} U_\text{eff} \cos\varphi$, is constant over time. The second varies periodically with time, at the circular frequency 2ω.

Here is an example which should already be familiar: An AC circuit contains a coil of inductance L. During the first quarter of an oscillation period, a magnetic field is produced in the coil. During the second quarter-period, the magnetic field again decays, and its energy $\frac{1}{2}LI_0^2$ is given back to the AC current source. During the third and fourth quarter-periods, the same cycle repeats itself but with the opposite signs for the current and the magnetic field. This second part leads to zero power summed over the full oscillation period.

It is therefore usual to distinguish two components of the current, known as the *active current*, with the amplitude $I_0 \cos\varphi$, and the idle current (or *reactive current*), with the amplitude $I_0 \sin\varphi$.C10.11 The ratio

C10.11. The reactive (idle) current has a phase shift of 90° relative to the voltage; the active current is in phase with the voltage. An application is described also in Sect. 13.11.

$$\frac{\text{Active current}}{\text{Reactive current}} = \frac{1}{\tan\varphi} \qquad (10.38)$$

is called the *loss factor* of this circuit. An AC generator must be able to *produce* a constant active power and simultaneously to "*lend*" a reactive power during every second quarter-period. Reactive power does not produce losses in this view, but it requires a large amount of "operating capital".

10.9 Transformers and Inductances (Chokes)

Knowledge of the self-inductance as a form of inertia opens up for us an understanding of the important topic of *transformers* or current and voltage converters for alternating current.

A transformer consists of two coils which encompass the same magnetic field (Fig. 10.21). One of them, the field or *primary* coil, has N_p turns in its windings. Its ends are connected to the AC current source. Its DC or OHMic resistance is negligible. Then the inductive voltage $U_{\text{L},0} = I_0\omega L$ acts between its terminals (Eq. (10.18)). The magnetic field which is associated with the current I passes not only through the primary coil, but also through the other coil, the induction or *secondary* coil, and in the N_s turns of that coil's windings, the secondary voltage U_s is induced. With the same magnetic field in both coils and no-load operation, according to the law of induction

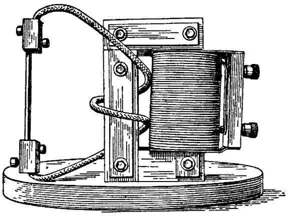


Figure 10.21 A current transformer for producing large currents

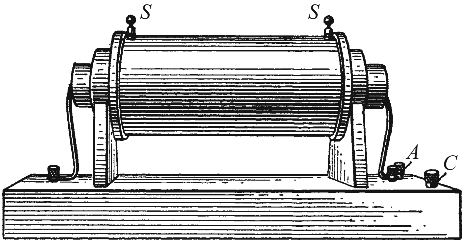


Figure 10.22 An induction coil

the two voltages have the same ratio as the numbers of turns of the coils, i.e. the *transformer equation* holds (**Exercise 10.13**):

$$U_{\mathrm{s},0}/U_{\mathrm{p},0} = N_\mathrm{s}/N_\mathrm{p}\,. \tag{10.39}$$

Therefore, by the proper choice of $N_\mathrm{s}/N_\mathrm{p}$, i.e. by choosing the transfer ratio, one can obtain almost any arbitrary increase or decrease of the voltage amplitudes. Transformers providing many hundred kV are operated today for a variety of research and technical purposes. Their main application, however, is in today's *long-distance power transmission*, which would be unthinkable without multiple voltage conversions. The consumer should receive voltages of at most a few hundred volts; apart from gross negligence, such voltages are not dangerous to life. The long-distance transmission lines, in contrast, must transport the electrical energy at high voltages and relatively low currents (e.g. 10^4 kW at 10^5 V and 10^2 A). Otherwise, the cross-sections of the wires would have to be much too large and the transmission lines would therefore be too heavy and too expensive.

Voltage reduction in the secondary circuit is accompanied by an increase in current strength. *Low-voltage transformers* are constructed with only a few turns in their secondary windings (e.g. 2 as shown in Fig. 10.21). They can be used to routinely provide currents of several thousand ampere for demonstration experiments. Technologically, this principle is used to construct *induction furnaces* for melting steel etc. The secondary circuit in this case consists of only a single loop. It can take the form of a circular trough lined with high-temperature resistant stone. The metal to be melted is placed in the trough. The induced currents may reach several tens of kA.

A special type of transformers are called "induction coils". Their primary coils (terminals A and C in Fig. 10.22) and secondary coils (terminals S) are coaxial and contain an iron core (not as a closed magnetic circuit); an example is shown in Fig. 10.22. Figure 10.23 shows a spark discharge from such a coil.

In the usual AC transformers, the necessary periodic variation of the magnetic field is produced by the AC current in the primary circuit. Induction coils instead operate with a *periodically interrupted direct current*. The periodic interruption can be provided by one of the numerous automatic switches developed for the purpose. The simplest types make use of *toggle*

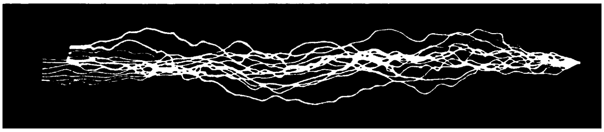


Figure 10.23 The spark discharge from an induction coil, about 40 cm long between the terminals (S in Fig. 10.22). The coil is equipped with a mechanical interrupter, and the exposure time for this photo was 1 second.

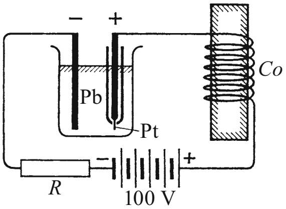


Figure 10.24 The production of toggle switching using an electrolytic interrupter, after A. WEHNELT. A positive electrode made of a 1 mm thick and 10 mm long platinum wire is mounted at the end of a glass nozzle in a dilute sulfuric acid solution. When a current flows, the wire is heated until it glows and surrounds itself with an insulating layer of gas, interrupting the current flow. This induces a voltage impulse in the coil Co, which destroys the gas layer, etc. The frequency of the toggle switching (with a fixed coil inductance) can be adjusted over a wide frequency range by varying the resistance R.

action. As an example, we mention the arrangement described in Vol. 1, Fig. 11.5 (WAGNER's hammer or vibrator, using the principle of a doorbell). A notable development is the production of toggle switching without moving parts (Fig. 10.24):

Exercises

10.1 Calculate the inductance L of one of the coils used in Video 6.1 as a "magnetic voltage meter". Their characteristics are: Number of turns in the windings $N = 4300$, length $l = 40$ cm and diameter $2r = 11$ cm. (Sect. 10.1)

10.2 A wire is bent into a circular ring. Its inductance is L_1. Determine the inductance L_N if N similar rings were collected into a bundle and all of them were carrying the same current I. (Sect. 10.1)

10.3 The coil in Fig. 10.7 has an inductance of $L = 50$ H. How large must one make the resistance R so that the current reaches one-half of its maximum value in a time of 10 s after closing the switch? (Sect. 10.2)

10.4 An AC voltage with a frequency of $\nu = 100\,\text{Hz}$ and an effective value of $U_{\text{eff}} = 150\,\text{V}$ is connected to a coil with the inductance $L = 0.3\,\text{H}$. Find the effective value of the resulting current I_{eff}. (Sect. 10.4)

10.5 An AC generator is connected to a series circuit with a resistor of $R = 13.7\,\Omega$ and a coil of inductance $L = 50\,\text{mH}$. The effective value of the current is found to be $I_{\text{eff}} = 12\,\text{A}$. Determine the frequency ν of the alternating current. (Sect. 10.4)

10.6 An AC generator operating at a frequency of $\nu = 50\,\text{Hz}$ and producing a voltage of effective value $U_{\text{eff}} = 80\,\text{V}$ is connected to a circuit with a resistance of $R = 10\,\Omega$. In order to limit the current to an effective value of $I_{\text{eff}} = 2\,\text{A}$, a coil of inductance L is connected in series with the resistor. Find the required value of L. (Sect. 10.4)

10.7 In an AC circuit, a resistor with the value $R = 10\,\Omega$ and a coil of inductance $L = 100\,\text{mH}$ are connected in series. What value must the frequency ν have, in order that an increase in L by 1 % would lead to an increase in the impedance by the same amount as a decrease in R by 50 % would decrease it? (Sect. 10.4)

10.8 An AC current of frequency $\nu = 50\,\text{Hz}$ is flowing through a copper wire with an OHMic resistance of $R = 5\,\Omega$ and a negligible inductance. By winding it into a coil, the impedance of the wire is to be increased tenfold. What inductance L must the resulting coil have? (Sect. 10.4)

10.9 A coil of inductance L and an OHMic resistance R is connected in series with a condenser of capacitance C (see Fig. 10.14). What is the relation between L and C that must hold in order for the impedance Z of the circuit to be equal to R, i.e. the current and voltage in the circuit are in phase? The frequency is $\nu = 1\,\text{kHz}$. (Sect. 10.6)

10.10 In this and the following two exercises, we investigate the experiment described in Figs. 10.19 and 10.20. The AC current source which is connected to the parallel circuit has an output voltage of $U = U_0 \sin \omega t$ with $U_0 = 73\,\text{V}$. The frequency ν and the values of C, L and R are given in Fig. 10.19.

a) Find the voltages $U_{\text{L},0}$ and $U_{\text{R},0}$ in the left-hand branch of the circuit, and use a phasor diagram (Fig. 10.12) to find the impedance Z_{RL} and the phase angle φ_1 between the current I_{RL} and the voltage U. The definition of φ is given in Eqns. (10.23) and (10.24).

b) Determine the current amplitudes $I_{\text{RL},0}$ and $I_{\text{C},0}$ in both branches of the parallel circuit. Draw these amplitudes and their phase angles relative to the applied voltage amplitude U_0 in a phasor diagram (the vector representing U_0 can be drawn in pointing to the right and parallel to the x-axis). From this diagram, find the amplitude I_0 and its phase angle φ_2 relative to U_0.

c) Compare the impedance found above and the phase angle φ_2 to the values found using Eqns. (10.33) and (10.34). (Sect. 10.7)

10.11

Find the logarithmic decrement Λ (sect. 11.7) of the parallel circuit in Fig. 10.19, and compare the result to the value given in Fig. 10.20. (Sect. 10.7)

10.12

a) From the value of I_0 found in Exercise 10.10, find the active current $I_0 \cos 15°$, that is the component of the current which is in phase with the voltage U, and the reactive current $I_0 \sin 15°$, the component which is out of phase with the voltage by 90°. b) Determine the power $\dot{W}$ delivered by the current source and compare it to the loss power $\dot{W}_{\mathrm{RLC}}$ in the parallel circuit. (Sect. 10.8)

10.13

For a pair of coils as shown on the left in Fig. 5.5, derive the transformer equation (10.39). An AC current source with output voltage $U_\mathrm{p} = U_{\mathrm{p},0} \cos \omega t$ is connected to the primary coil, and an AC voltmeter is attached to the secondary coil. The primary coil has N_p turns in its windings, its cross-sectional area is A_p, and its length is l_p. The secondary coil, which surrounds the primary coil on its outside, has windings with N_s turns. The OHMic resistance in the primary circuit can be neglected, and that of the voltmeter is infinite. Derive the ratio of the secondary voltage $U_{\mathrm{s},0}$ to the primary voltage $U_{\mathrm{p},0}$ by making use of the law of induction and knowledge of the inductance of the primary coil (since the two voltages have opposite signs, we mean here the ratio of their magnitudes). (Sect. 10.9)

10.14

a) Find the time-averaged value of the power $\overline{\dot{W}} = (1/T) \int \dot{W} \,\mathrm{d}t$ which is delivered by the current source in Exercise 10.13.

b) How does this power change if the voltmeter in the secondary circuit is replaced by a resistor with a finite OHMic resistance R? A qualitative answer will suffice here. (Sect. 10.9)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_10) contains supplementary material, which is available to authorized users.

Electrical Oscillations 11

11.1 Preliminary Remark

In the previous chapter, we dealt with alternating currents, including circuits containing a condenser and a coil ("tank circuits"). In the present chapter, we will identify these circuits as *oscillators* and investigate their properties.

11.2 Free Electrical Oscillations

In Fig. 10.14, we saw a series circuit, and in Fig. 10.18 a parallel circuit. In both of these figures, the *AC generator* was simply indicated by the symbol $\sim$.

An AC generator can be fabricated from a DC current source and a variable resistor. This is accomplished using the scheme shown in Fig. 11.1: Periodic variations of a resistance E or D around a mean value U/I produce an alternating current, superposed on a direct current. This yields an AC voltage between the terminals a and b.

With these setups, containing so-called "tank circuits" (at the right in each figure), we can repeat the experiments shown in Figs. 10.14 and 10.18 at very low frequencies (ν of order 1 Hz). We require

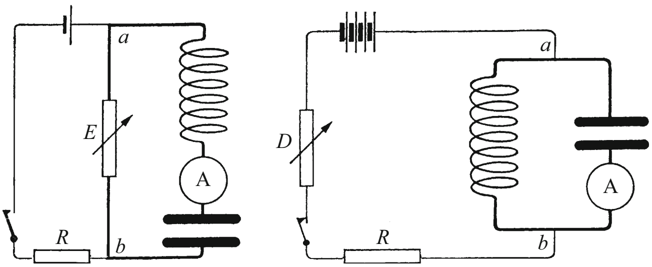


Figure 11.1 A series LC circuit and a parallel LC circuit ("tank circuits") as oscillators, each connected to an AC generator. E and D are periodically variable resistors, while R is a fixed resistor. The switches are used to produce damped oscillations (for $\nu = 1\,\text{Hz}$, the resistance E in the left-hand image must have a value of the order of $100\,\Omega$, and D, in the right-hand image, must be around $10^4\,\Omega$. The current source at the left is a storage battery (2 V), while on the right, around 100 V is required).

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_11

a coil with a very large inductance ($L \approx 10^3$ H) and a condenser with a large capacitance (C up to 50 µF). With the switch closed, a sliding contact on the resistors is moved periodically back and forth, either by hand or using an eccentric and connecting rod driven by a motor.

In Fig. 11.1, at left, the resistor E has a low resistance, while at the right, D has a large value. This matches the internal resistance of the "generator" to the resistance U/I of the circuit between the terminals a and b (where it is attached to the generator). The resistance of the series circuit may be rather low (Fig. 10.17), while it is high for the parallel circuit (Fig. 10.20).

The AC current flowing through the coil and the condenser (heavy lines), which has a constant amplitude, can be observed with an ammeter of short response time (0.12 s), or with an oscilloscope. For both the series circuit and the parallel circuit, we find the largest amplitudes around $\nu \approx 1$ Hz.

Then we see something new: We leave the resistances E and D constant and simply open or close the switch. Then we can observe an alternating current in both the series circuit and the parallel circuit, with a *decaying* amplitude; or, put differently, we observe damped oscillations. *Therefore, both series circuits and parallel circuits are* oscillators. They each contain a storage device for electrical energy (the condenser) and for magnetic energy (the coil). Closing or opening the switch is sufficient to produce damped oscillations (i.e. alternating currents with decaying amplitudes) by impulse excitation in these *oscillator circuits*.

The oscillations start with deflections in opposite directions, depending on whether the switch is closed or opened. In the ideal limiting case, an electrical oscillator circuit would oscillate without losses at a constant amplitude following an impulse excitation. In this limit, series and parallel circuits are equivalent. For this limiting case, there is a clear-cut mechanical analog (Fig. 11.2). Today, it can be found in most school books on physics and needs no further explanation.

"For electrical oscillations, there is a clear-cut mechanical analog (Fig. 11.2). Today, it can be found in most school books on physics and needs no further explanation".

The frequency of these electrical oscillations is the *resonance frequency* which we have already encountered in Sect. 10.6:

$$\nu_0 = \frac{1}{2\pi\sqrt{LC}} \tag{10.32}$$

(in the parallel circuit, this holds only when $R^2C/L \ll 1$; see Eq. (10.35)),

i.e. the frequency at which the inductive resistance of the coil, ωL, is just as large as the capacitive resistance of the condenser, $1/(\omega C)$ in the series circuit at resonance.

In order to obtain electrical oscillations of higher frequency by impulse excitation of a series circuit (Fig. 11.1), we would have to make L and C much smaller, as can seen from Eq. (10.32). Then the energy initially stored in the condenser, $W_e = \frac{1}{2}CU^2$ (Eq. (3.20)), is

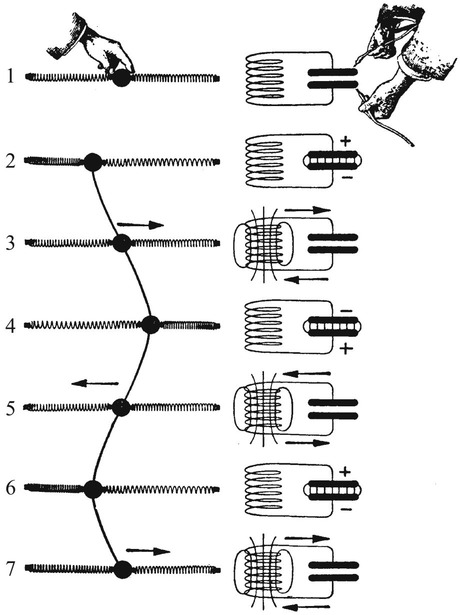


Figure 11.2 The periodic exchange of potential and kinetic energy in mechanical oscillations (left), and of electrical and magnetic energy in electrical oscillations (electrical oscillator circuit, right). The arrows on the right mark the direction of motion of the electrons within the circuit.

also rather small. As a result, we would then have to use higher voltages; that however causes an annoying problem: At high voltages, an arc jumps across the gap between the switch contacts even before the switch is closed. *One has to live with these disturbing sparks.*C11.1 *But they can be utilized in a useful way*, namely:

1. as a periodically-operating automatic switching device, and
2. as an ammeter with a very short response time.

The spark acts as an *automatic switch* for example in Fig. 11.3. Instead of a moving and a fixed contact, we see there a *spark gap* consisting of two metal balls. Two thin wires serve to charge the condenser from some sort of current source, e.g. an influence machine. When a certain maximum voltage U is reached, the spark jumps over the gap and starts the oscillator circuit. This operating voltage U can be adjusted by changing the distance between the balls of the spark gap.

The spark can also be employed as an *ammeter with very short response time*, owing to the dependence of the light density that it emits

C11.1. Using a modern pulse generator and a sensitive oscilloscope, such sparks can easily be avoided. Nevertheless, the example that follows here is very useful, since it illustrates the early precursors of "wireless telegraphy", of enormous importance today in its modern forms.

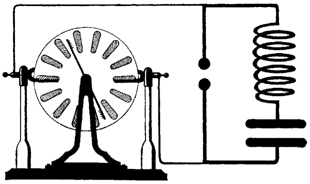


Figure 11.3 A spark gap as the switch for an electrical series circuit as in Fig. 11.1. The resistors R and E have been removed, and only the switch remains.

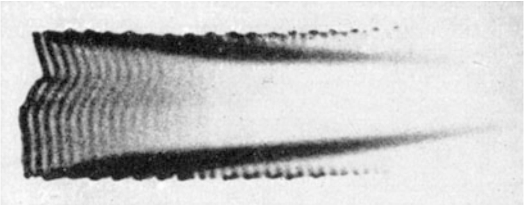


Figure 11.4 The detection of electrical oscillations from a spark (its duration is of the order of 10^{-3} s; a “FEDDERSEN spark”, 1859.[C11.2] The image is the negative of a photograph taken by B. WALTER).[C11.3]

C11.2. Martin Henke: “*Flinke Funken im schnellen Spiegel – Berend Wilhelm Feddersen und der Nachweis der elektrischen Schwingungen*”, Dissertation, University of Hamburg, 2000. This document contains a detailed account of FEDDERSEN’s discovery.

C11.3. Similar photographs of lightning discharges can be found in: B. Walter, *Annalen der Physik*, 4th series, **21**, 223 (1906).

on the current. The light density exhibits two maxima during each oscillation cycle. To make these variations in light density visible, images of the spark which follow each other closely in time must be separated spatially. A rapidly rotating polygonal mirror provides a simple method of doing this. In Fig. 11.4, photographs of sparks taken with this method are shown. The frequency of the oscillations was 50 kHz. Initially, the periodic variations in the light density can be readily distinguished; in later images, they are obscured by clouds of glowing metal vapor. In the early images, one can also recognize the direction of the current during the individual maxima. The brighter end of the spark always indicates the negative pole.

The high voltages which accompany this excitation of high-frequency electrical oscillations can be used for impressive demonstration experiments. These are discussed in the following section.

11.3 High-Frequency Alternating Currents for Demonstration Experiments

1. The TESLA coil.

In Fig. 11.5, the coil *Co* of a high-frequency electrical oscillator circuit serves at the same time as the primary winding of a transformer. It consists of only a few turns, e.g. $N_\mathrm{p} \approx 3$. At high frequencies, it still has a large inductive resistance $U_0/I_0 = \omega L$. As a result, we can produce voltage differences of several 10^4 V between its ends without increasing the current to more than a few ampere. The number of turns N_s of the secondary-coil windings is considerably greater, e.g. $N_\mathrm{s} \approx$ several hundred turns. Therefore, voltages of several 10^5 V can be readily produced at the ends of the secondary coil. There, long, reddish-blue forked sparks jump between the terminals (Fig. 11.5). Often, one end of the secondary or induction coil is grounded (to a water pipe or something similar). The free terminal is then the source of lively bundles of reddish, branched sparks

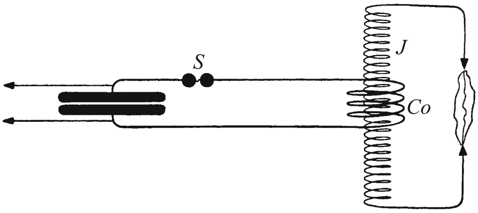
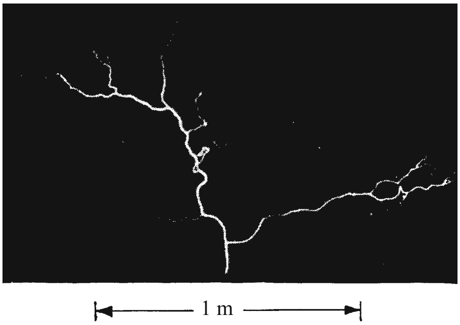


Figure 11.5 A TESLA transformer. The primary coil *Co* is the coil of a damped electrical oscillator and carries a high-frequency AC current ($\nu \approx$ 400 kHz; the arrows indicate leads to the current source, e.g. a resonant transformer). The secondary coil of the resonant transformer also makes up an oscillator circuit together with the condenser *C*; its frequency is the same as that of the line voltage, usually 50–60 Hz (current resonance, see Sect. 10.7). (**Video 11.1**) (**Exercise 11.3**)

Figure 11.6 Snapshot image (0.01 s) of the brush discharge from the terminal of a TESLA coil (Fig. 11.5)



Video 11.1:
“TESLA Coil”
http://tiny.cc/adggoy
The video shows experiments with an historic apparatus from the early 20th century, which is still used for demonstrations today at Cornell University.

(“brush discharges”), which weave around and can be over a meter long (Fig. 11.6). Their physiological harmlessness is surprising.[C11.4]

2. *Demonstration of self-inductance in non-coiled conductors.* For alternating current, the inductive resistance $U_0/I_0 = \omega L$ of a conductor is in general large compared to its OHMic resistance *R*. With high-frequency AC, this effect can be demonstrated using a “coil” consisting of a single turn: a wire loop.

In Fig. 11.7, a thick copper-wire loop is carrying high-frequency alternating current in the primary circuit of a TESLA coil. It is bridged at its center by an incandescent light bulb, which would be practically short-circuited for direct current. Nevertheless, the bulb glows brightly. The ratio U/I, defined as the resistance, must therefore be much greater for the high-frequency AC than for DC. This simple experiment shows the inertia of the magnetic field in an obvious way. It contains nothing essentially new; but it is important, since the beginner tends to forget about self-inductance in conductors which are not wound into a coil (**Exercise 11.2**).

3. *The skin effect.* We can imagine a wire to be composed of a thin central axis and concentric, tube-shaped layers surrounding it. *Its inductance L is smaller within the outer layers than within the inner layers.*

C11.4. This is explained in **Video 11.1** at 9:40 min. See also e.g. Christopher Gerekos, www.teslacoildesign.com/docs/TheTeslaCoil-Gerekos.pdf or also https://en.wikipedia.org/wiki/Tesla_coil. Compare the last section of the latter reference. While there are no electric shocks and no electrolysis effects of the high-frequency current, there are still health hazards, as explained there.

“...the beginner tends to forget about self-inductance in conductors which are not wound into a coil”.

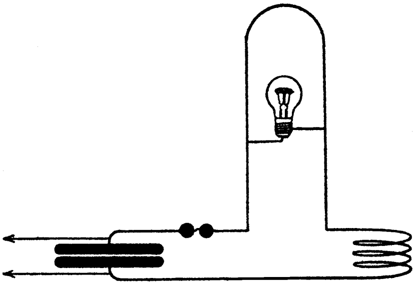


Figure 11.7 A demonstration of the inductive resistance of a wire loop

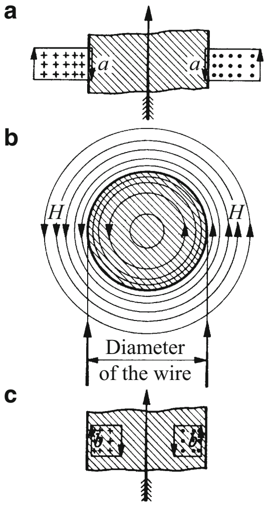


Figure 11.8 Magnetic field lines around and within a wire (the wire is indicated by shading), and their inductive effects (the feathered arrow shows the conventional direction of the electric current; in part b, it points out of the plane of the page towards the reader)

Justification: In Fig. 11.8, we see in part b a straight current-carrying conductor in cross-section (shaded). The conductor is surrounded in the usual manner by the ring-shaped, closed-loop field lines of a magnetic field H. These field lines however not only surround the conductor on its exterior, but also penetrate into its interior. Each of the tube-shaped layers which is carrying a current must indeed be surrounded by magnetic field lines.[C11.5] Several of them are sketched in Fig. 11.8.

C11.5. See Eq. (6.13), which follows from MAXWELL's equation (6.23). These equations were in fact derived for empty space, but as long as the current conductor in Fig. 11.8 is not ferromagnetic, the difference between the conducting material and vacuum can be neglected (see Chap. 14).

In addition, one section of the conductor is shown twice as a longitudinal section (Fig. 11.8, parts a and c). In both images, the direction of the current is indicated by a long feathered arrow. Furthermore, the magnetic field lines are shown where they penetrate the plane of the image (as • or +). Thus, in part a, above, we see the points of penetration of several *exterior* magnetic field lines, while below, in part c, we see those of several *interior* magnetic field lines. Variations of this magnetic field with time induce an electric field with closed-loop field lines. Pairs of these field lines are drawn in parts a and c as rectangles, for the case that the current is *increasing*. At the surface of the wire, these newly-formed electric fields due to self-inductance have opposite directions. The arrows at a point downwards, while the arrows at b point upwards; therefore, the induced electric fields cancel each other to a great extent. Along the central axis of the wire, in contrast, this compensation is lacking. There, the induced field is opposite to the external applied field (feathered arrow); as a result, the induced field hampers the increase of the current. When the current is decreasing, the opposite occurs *within* the wire: along the inner axis of

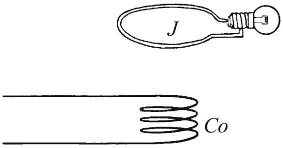


Figure 11.9 Induction with high-frequency alternating currents (coil *Co* as in Fig. 11.5)

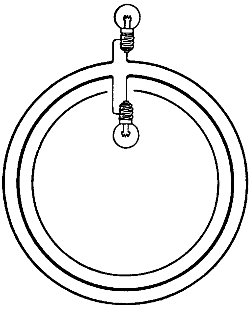


Figure 11.10 Demonstrating the skin effect

> the wire, the induced field and the applied field have the same direction, so that the induced field hinders the decrease of the current. Result: Along the central axis of the wire, in contrast to its surface, its self-inductance is stronger.

This uneven distribution of induction within the cross-section of the wire becomes especially noticeable with alternating currents at high frequencies. To detect this current displacement ("skin effect"), we use the setup sketched in Fig. 11.9. The coil *Co* carries a high-frequency alternating current. It induces currents in the induction coil *J*, a thick copper-wire ring. An incandescent lamp included in the circuit is used to estimate the current strength. We then surround the copper ring with a concentric copper tube (Fig. 11.10). The walls of the tube have the same cross-sectional area as the wire. A similar lamp is connected between the ends of the tube and of the copper wire loop. These two *induction coils*, one inside the other, are now brought near the field coil *Co* in Fig. 11.9. The lamp connected to the ends of the tube shines brightly, while the lamp between the ends of the wire glows only a dull red or not at all (see **Video 11.1**).

4. The *detection of closed-loop electric field lines*. According to the detailed explanation of induction processes (Sect. 6.1), there should be closed-loop electric field lines resulting from induction. They could unfortunately not be made visible using insulating powder (e.g. fine gypsum crystals). With the high-frequency AC currents in electric oscillator circuits, we can return to this topic and fill in this missing link in our earlier discussion by making the closed-loop field lines visible in a clear-cut way.

The apparatus is shown in Fig. 11.11. The field coil *Co*, with just one turn, produces a high-frequency AC magnetic field. Its field lines are perpendicular to the plane of the page in the figure. This rapidly varying magnetic field should surround itself, according to Fig. 6.2, with closed-loop electric field lines.

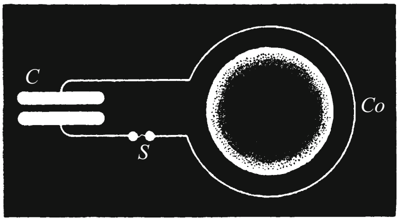


Figure 11.11 The detection of closed-loop electric field lines, the "electrodeless ring currents". Noble gases at low pressure, for example neon, already begin to glow at electric field strengths of ca. 20 V/cm.

Now we bring up a glass bulb filled with neon at low pressure into the region of these closed-loop electric field lines; a ring-shaped region within the bulb glows brightly, widely visible to a large audience. We see an image, albeit rough, of the AC electric field with its closed-loop field lines, with no beginning or end. Knowledge of these field lines is essential for the elucidation of electromagnetic waves in later chapters. This demonstration is therefore quite important for our understanding (**Video 11.1**).[C11.6]

C11.6. ...Even though it does not prove the existence of closed-loop electric field lines, as is emphasized in **Video 11.1** at 5:10 min.

11.4 The Production of Undamped Electrical Oscillations Using Feedback with Triodes

The electrical oscillations discussed so far, which were produced by impulse excitation, were all *damped*. Energy losses caused their amplitudes to decrease with time. A damped oscillation has more than one frequency (Vol. 1, Sect. 11.4); not only its eigenfrequency ν_0, but also a broad frequency *range*, a continuous frequency spectrum, occurs within the damped oscillations. This is rather disturbing for many physical and technical purposes. One often requires alternating currents with constant amplitudes over longer times, and with a well-defined frequency. Therefore, it is very useful to be able to produce *undamped* electrical oscillations.

In mechanics, the corresponding task was accomplished long ago (Vol. 1, Sect. 11.2) by making use of the technique of *autocontrol or positive feedback*. Pendulum clocks and all kinds of watches are well-known examples. The control of an electrical oscillator by feedback will be treated by first considering oscillations at very low frequencies ($\nu \approx 1\,\text{Hz}$). We use the circuit already described in Fig. 11.1 (right-hand image); however, with a small modification as shown in Fig. 11.12 at the left: The "AC generator" again consists of a DC source and a variable resistor R_s, but it is connected only to a part of the coil between the points a and b.

Repeated, uniform variations of R_s at the rhythm of the resonance frequency ν_0 of the oscillator circuit produce an alternating current of constant amplitude. In order to achieve *autocontrol*, the circuit itself must produce this periodic variation in R_s. This can be accomplished

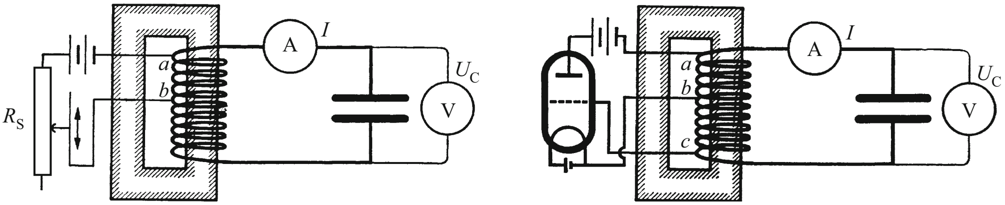


Figure 11.12 The production of undamped electrical oscillations in a parallel tank circuit at very low frequencies ($\nu \approx 1\,\text{Hz}$); at the left using *external control*, at the right with *autocontrol* (feedback) ($R_s \approx 10\,\text{k}\Omega$, $C = 50\,\mu\text{F}$, $L = 10^3\,\text{H}$).

for example by using a *triode* (electron tube with three electrodes) as a "switch" or "valve" (Fig. 11.12, right side).

In a triode, within the evacuated bulb between the thermal-emission cathode and the anode plate, there is a control grid. The resistance of a vacuum tube of this type is of order of $10\,\text{k}\Omega$. It can be varied periodically over a wide range by applying an oscillating electric field between the cathode and the grid[1]. The required voltage can be obtained for example by induction from the section $b - c$ of the coil. In this way, the oscillator circuit at the right in Fig. 11.12 produces an undamped AC output at a frequency $\nu \approx 1\,\text{Hz}$. The time dependence of the current and the voltage as well as the phase difference between them can be registered by rotating-coil galvanometers with a short response time or with an oscilloscope.

This method of autocontrol (often called *feedback*) can be applied with common electron tubes up to frequencies of around 100 MHz. Figure 11.13 shows an example for frequencies of a few hundred kHz. A small lamp is used as an indicator for the AC current.

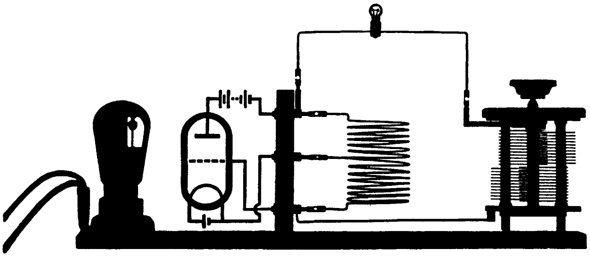


Figure 11.13 The production of undamped electrical oscillations with frequencies of the order of 500 kHz. This figure and some of the following figures show shadow projections of operable setups. The triode can be seen at the far left. The circuit diagram is drawn on a glass plate (cf. Fig 11.12, right-hand image).

[1] In Figs. 11.12, 11.13 and 11.17, imagine that a 1.5 volt battery is inserted into the wire leading to the grid of the triode. POHL gave details in his "*Elektrizitätslehre*", 21st edition, Section 17.5. English: See e.g. https://zipcon.net/~swhite/docs/physics/electronics/Valves.html.

Today, *vacuum-tube* triodes have been almost completely replaced by *solid-state* triodes (transistors).[C11.7]

C11.7. In Chap. 27, Sect. 3 in the 21st edition of Pohl's "*Elektrizitätslehre*", the operation of a *vacuum-tube* triode was explained in terms of the first experimental *solid-state* triode, which had not yet been developed sufficiently for technical applications (R. Hilsch and R. W. Pohl, *Zeitschrift für Physik* **111**, 399 (1938); see also **Video 1**, 23:30 min, http://tiny.cc/z8fgoy).

11.5 Feedback with Diodes

Autocontrol with diodes also begins with the discussion in Sect. 11.2. Here again, an AC generator is used in order to maintain electrical oscillations with a constant amplitude. As generator, again the combination of a DC current source with a periodically-variable resistor is employed. The variation of the resistance in this case is the result of a particular property of the resistor employed. Its I-V characteristic (i.e. the dependence of the voltage drop across the resistor on the current it is carrying) exhibits the form shown for two examples in Fig. 11.14: In some range, the differential resistance $\mathrm{d}U/\mathrm{d}I$ becomes negative! Conductors with this characteristic are referred to in the following as *diodes*.

How diodes can effect the autocontrol of electrical oscillations is explained using two demonstration experiments as examples (Fig. 11.15). Both make use of *diodes* whose properties change *audibly* during the periodic control process. In a series circuit, the *electric arc E* represents a diode of type S; in a parallel circuit, a small Wehnelt *interrupter D* (Fig. 10.24) acts as an N-type diode. In the electric arc, the alternating current from the series circuit adds to the direct current from the current source. This changes the current in the arc periodically, and along with it (due to heating), the volume of the arc: The arc "sings". In the Wehnelt interrupter, the alternating voltage between the terminals a and b of the parallel circuit adds to the direct voltage from the current source. This causes a periodic variation of the voltage between the electrodes of the diode D, and with it, the gas bubble surrounding the glowing platinum wire expands or contracts: The gas bubble "sings". In both tank circuits, the

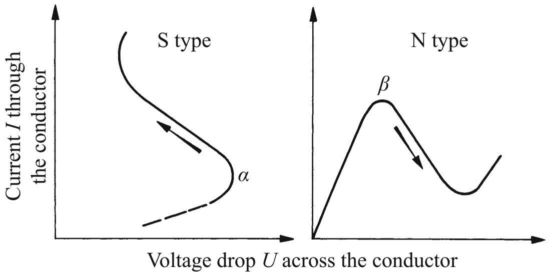


Figure 11.14 I–V characteristic curves, referred to as S and N types. They are found for conductors with a negative differential resistance $\mathrm{d}U/\mathrm{d}I$ (schematic drawing). *Initially*, in the S type, the current increases up to a voltage maximum α; in the N type, the voltage increases up to a current maximum β.

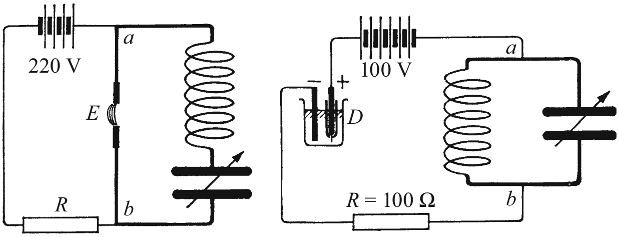


Figure 11.15 Autocontrol (feedback) of electrical oscillator circuits using diodes whose properties change *audibly* ($\nu \approx 1\,\text{kHz}$). At the left is a *series* circuit similar to the schematic in Fig. 11.1 (left-hand image); here, E is an electric arc. At the right is a parallel tank circuit similar to the schematic in Fig. 11.1 (right-hand image); here, D is a small WEHNELT interrupter (the coil is similar to Fig. 10.2, right, but without an iron core: $L = 3.3\,\text{mH}$; the condenser has $C \approx 1\,\mu\text{F}$. The electric arc is struck between two pure carbon electrodes of 1 cm diameter).

frequency of the tone can be adjusted by varying the capacitance C of the condenser.

> In the feedback circuit using a parallel tank circuit, one can leave off the condenser altogether. Then only a *toggle oscillation* at an audible frequency remains (Vol. 1, Sect. 11.17); it can be varied by changing the value of the resistor R (cf. Fig. 10.24). This once again demonstrates the transition from harmonic oscillations to toggle oscillations.

A different conductor with an N-type characteristic (Fig. 11.14, right side) is the Dynatron (cf. 21st German edition of this book, Sect. 17.5; see also https://en.wikipedia.org/wiki/Dynatron_oscillator), as well as semiconductor diodes (e.g. tunnel diodes). With these, oscillator circuits for frequencies up to 100 GHz can be assembled.

11.6 Forced Electrical Oscillations

Let us consider some sort of mechanical pendulum, which swings following an "impulse excitation" or undamped with "feedback" at its resonance frequency, ν_0. But every pendulum can also be *forced* to oscillate at a different frequency by a suitable *external excitation source*, so that the pendulum oscillates as a *resonator*. To accomplish this, we let a periodic force or torque act with the desired frequency on the pendulum. This process of *forced oscillations* was treated in some detail in Vol. 1 (Sect. 11.10) because of its general importance.

Similar conclusions hold for forced electrical oscillations. Instead of a torsional pendulum with a flywheel and a spiral spring, we consider an electrical oscillator circuit with a coil and a condenser; instead of a periodically recurring *torque*, we require a periodically varying *voltage*. The latter can be produced *without a conducting connection*

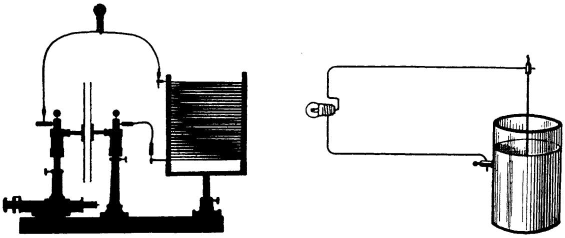


Figure 11.16 Two oscillator circuits ("tank circuits") with which high-frequency AC currents can be demonstrated by means of resonant excitation. A suitable excitation source is the circuit shown in Fig. 11.13. The resonance frequency of the left-hand circuit can be adjusted using the micrometer screw on the parallel-plate condenser; that of the right-hand resonator is fixed.

between the excitation source and the resonator. We start with two qualitative examples:

1. We use the two tank circuits shown in Fig. 11.16 as resonators; the excitation source could be the circuit in Fig. 11.13, which oscillates continuously, without damping. The resonators are set up near the excitation source, so that the magnetic field of its coil can induce a voltage in the windings of the resonator coil. By adjusting the condenser in the tank circuit, we can readily obtain the resonant frequency. Then the lamp in the resonator circuit shines brightly. Both resonator circuits have a small damping, so that their resonance curves are rather sharp.

2. High-frequency circuits with undamped oscillations often lack physical clarity. Coils and condensers can no longer be considered separately as discrete devices; often, the electrodes of vacuum tubes already provide the necessary capacitance C. One case of this kind can be seen in Fig. 11.17, at the left; it is an oscillator circuit with a resonance frequency of 1 MHz. We see only a coil which is tapped on one side, and a vacuum tube. At the right in the figure, in contrast, we see a regular oscillator circuit, containing a coil with two turns and a rotary variable condenser. The left-hand circuit serves as the *excitation source*, while the right-hand circuit is the resonator. As an indicator for the forced oscillations, we again use a small light bulb. In this way, we can produce high-frequency AC currents in an oscillator without unnecessary technical accessories.

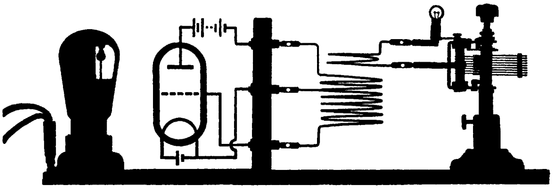


Figure 11.17 At the left: An oscillator circuit with autocontrol (feedback) for $\nu \approx 1\,\text{MHz}$. At the right: A tank circuit in which oscillations can be produced by resonance at the same frequency

11.7 A Quantitative Treatment of Forced Oscillations in a "Tank Circuit"

In the preceding chapter, in Figures 10.14 and 10.18, we varied the voltage between the terminals *a* and *b* of the oscillator circuits by using a generator which produces a sinusoidal alternating current. That is, we input an alternating current into the circuits. But now we say: The circuits are devices which can oscillate; we excite them to oscillations as *resonators*, and we employ an alternating current source for the excitation. We wish to compare these *forced electrical oscillations* with forced mechanical vibrations or oscillations like those that were discussed in Vol. 1, Sect. 11.10.

The alternating current in Fig. 10.14, that is in a *series tank circuit* containing a coil and a condenser, was described by Eq. (10.30). With this equation, we could first calculate the amplitude I_0 of the current for various frequencies ν (Fig. 11.18b), and then, by referring also to Eq. (10.29), we could obtain the amplitude $U_{\mathrm{C},0}$ of the voltage on the condenser (Fig. 11.18a). Furthermore, with Eq. (10.31), we could calculate the phase angle φ, and with it, the phase difference between the excitation voltage U and the condenser voltage U_C (Fig. 11.18c); and finally, we computed the average power $\dot{W} = \frac{1}{2}I_0^2 R$ dissipated by the resonator (Fig. 11.19); this curve holds at the same time for the average magnetic energy $W_\mathrm{m} = \frac{1}{2}(\frac{1}{2}LI_0^2)$ stored in the resonator. Its value can be read off the right-hand ordinate scale.

The values of L, C and R were chosen to be about the same as those for the demonstration experiment in Fig. 10.16.

The formal similarities between the forced oscillations of an electrical series-circuit oscillator and the forced oscillations of a mechanical harmonic oscillator are evident. The content of equations (10.30) and (10.31) is illustrated in a very clear-cut way by Fig. 11.18.

In the energy resonance curve (Fig. 11.19), the ratio of the resonance frequency of the resonator to its linewidth H is called the *quality factor* or *Q-value*:

$$Q = \frac{\nu_0}{H}\,. \tag{11.1}$$

This quantity serves to determine *experimentally* several important values which characterize the oscillator circuit, namely its *logarithmic decrement* Λ (Vol. 1, Sect. 11.10) and its *damping constant* $\Lambda\nu_0$. When the logarithmic decrement Λ is ≤ 1, then the following relation holds:

$$\frac{H}{\nu_0} = \frac{\Lambda}{\pi} = \frac{1}{Q} \tag{11.2}$$

($1/Q$ is also called the *loss factor*).

$1/\Lambda\nu_0 = 1/\pi H = \tau_r$ is the *relaxation time* within which the *amplitude* of the forced oscillations approaches the fraction $(1 - 1/e) \approx$ 63 % of its steady-state value.

We *calculate* these quantities, in the case of *weakly-damped series and parallel tank circuits*:C11.8

C11.8. A good introduction to the physics of vibrations and oscillations can be found for example in: H.J. Pain, "*The Physics of Vibrations and Waves*", John Wiley, 5th ed., 1999 (Chaps. 2 and 3).

$$\frac{\Lambda}{\pi} = R\sqrt{\frac{C}{L}}\,. \tag{11.3}$$

These quantities can also be obtained experimentally using the following relations:

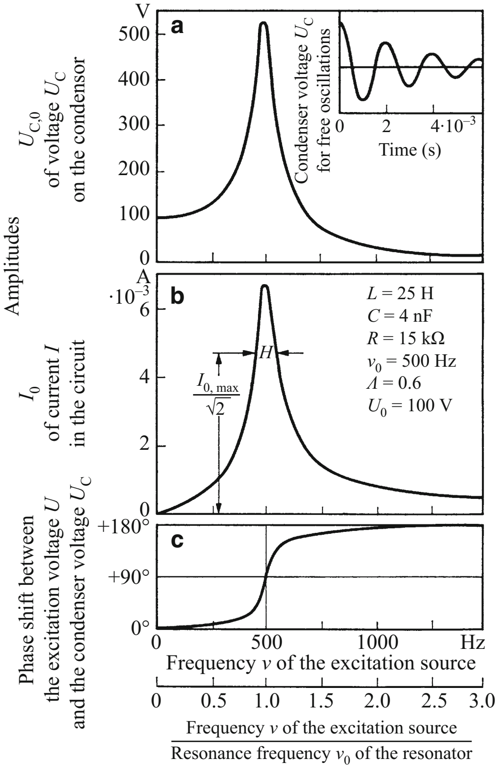


Figure 11.18 The characterization of forced electrical oscillations in a series tank circuit, calculated as in Sect. 10.6. The angle plotted in Part c is $(\varphi + 90°)$, where φ is defined as in Eq. (10.24). The linewidth (full width at half maximum, FWHM) H is the frequency range at whose limits the current (effective value or amplitude) has decreased to $1/\sqrt{2}$ of its maximum value. The maximum value of the current (Part b) lies as usual at ν_0.C11.9 With extremely strong damping, i.e. a logarithmic decrement of $\Lambda > 1$, the maximum deflection in Part a occurs neither at the resonance frequency ν_0 of the undamped oscillator, nor at the slightly lower resonance frequency of the freely oscillating damped oscillator circuit, but rather at the frequency $\nu = \nu_0\sqrt{1 - 0.5(\Lambda/\pi)^2}$.

C11.9. Note the analogy to the mechanical harmonic oscillator: The momentary deflection α and the amplitude α_0 of the torsional pendulum (Figs. 11.42a and b in Vol. 1) correspond in Fig. 11.18a to the condenser voltages U_C and $U_{C,0}$. The amplitude of the angular velocity $(\mathrm{d}\alpha/\mathrm{d}t)_0$ (Fig. 11.43 in Vol. 1) corresponds in Fig. 11.18b to the current amplitude, I_0.

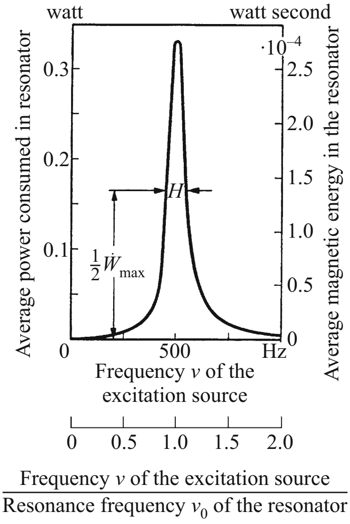


Figure 11.19 The energy resonance curve of forced electrical oscillations (series tank circuit). Even with extremely strong damping, i.e. $\Lambda > 1$, its maximum falls at ν_0, that is at the eigenfrequency of the undamped resonator. The FWHM H here is the frequency range at whose limits both the power dissipated by all damping mechanisms, as well as the average stored magnetic energy, have decreased to *half* their maximum values.

For the *series circuit*,

$$\frac{\Lambda}{\pi} = \frac{H}{\nu_0} = \frac{U_{\text{C},0} \text{ for } \nu = 0}{U_{\text{C},0} \text{ for } \nu = \nu_0} = \text{reciprocal of the resonant voltage increase} \tag{11.4}$$

and for the *parallel circuit*,

$$\frac{\Lambda}{\pi} = \frac{H}{\nu_0} = \frac{I_0 \text{ for } \nu = \nu_0}{I_{\text{L},0} \text{ for } \nu = \nu_0} = \text{reciprocal of the resonant current increase}\,. \tag{11.5}$$

Derivation: For the series circuit, at the frequency $\nu = 0$, $U_{\text{C},0}$ is equal to the amplitude U_0 of the power supply. At the resonance frequency, $\nu_0 = 1/(2\pi\sqrt{LC})$, we find from Eqns. (10.29) and (10.30) for the voltage amplitude $U_{\text{C},0}$ of the condenser

$$U_{\text{C},0} = U_0\sqrt{\frac{L}{C}} \cdot \frac{1}{R}\,,$$

and, making use of Eq. (11.3),

$$\frac{U_{\text{C},0} \text{ for } \nu = 0}{U_{\text{C},0} \text{ for } \nu = \nu_0} = R\sqrt{\frac{C}{L}} = \frac{\Lambda}{\pi}\,. \tag{11.4}$$

Parallel circuit: The amplitude I_0 of the current in the leads to the parallel circuit at the circular frequency $\omega_0 = 1/\sqrt{LC}$ follows from Eq. (10.33),

$$I_0 = \frac{U_0\omega_0 CR}{\sqrt{R^2 + L/C}}\,.$$

At the same circular frequency, the current amplitude $I_{\text{L},0}$ in the coil is given by Eq. (10.21),

$$I_{\text{L},0} = \frac{U_0}{\sqrt{R^2 + (\omega_0 L)^2}}\,.$$

C11.10. See e.g.: B. Kurrelmeyer and W.H. Mais, "*Electricity and Magnetism*", Van Nostrand, Princeton 1967, Chap. 14.

Thus, for $\nu = \nu_0$[C11.10], we have

$$\frac{I_0}{I_{L,0}} = R\sqrt{\frac{C}{L}} = \frac{\Lambda}{\pi}\,. \qquad (11.5)$$

In spite of the qualitatively different form of the current $I(\omega)$ in the parallel circuit (Eq. (10.33)), where the AC resistance has its maximum at the resonant frequency (see Fig. 10.20) and the current I is thus at its *minimum*, the relationship between Λ, Q, and the width H (which is however differently defined in this case) as given in Eq. (11.2) still holds here. H is defined here as the frequency range at whose limits the power consumed is twice as large as at the minimum (!); that is, the AC resistance has decreased by a factor of $1/\sqrt{2}$.

Exercises

11.1 In a loss-free LC circuit, at the time $t = 0$ the condenser has a charge of Q_0 (Fig. 11.2, right), and its voltage is thus $U_{C,0} = Q_0/C$. An electrical oscillation begins. Find the time-dependent voltage $U_C(t)$ on the condenser and the time-dependent current $I(t)$ in the circuit. (Sect. 11.2)

11.2 In an electrical oscillator circuit (Fig. 11.2), two coils with inductances of L_1 and L_2 are connected in series. Their "mutual inductance", i.e. the mutual influence of their magnetic fields, is presumed to be negligible (for example because they are sufficiently far apart; this is not essential to the principle of this exercise). The amplitude of the voltage on the condenser is $U_{C,0}$. Find the amplitude of the voltage $U_{L1,0}$ on the coil of inductance L_1. (Sects. 11.2, 11.3)

11.3 For the TESLA coil in Video 11.1 and Fig. 11.5, which we consider to be an oscillator circuit, the frequency ν_0 is to be estimated. In the experiments shown in Figs. 11.9–11.11, only the large primary coil is connected to the condenser. The condenser, a Leyden jar (Fig. 2.54), has a diameter of $d = 12\,\text{cm}$ and has metal-foil electrodes up to a height of $h = 18\,\text{cm}$ inside and outside. The thickness of its glass wall is $t = 1.52\,\text{mm}$, and we may assume a value of $\varepsilon = 7$ for its dielectric constant (Table 13.1). The coil consists of $N = 8$ circular windings with a radius of $a = 12.5\,\text{cm}$. The diameter of the wire is $2b = 3\,\text{mm}$. Under the assumption that $b \ll a$, the inductance of a single circular ring is given by $L = \mu_0 a\,(\ln 8a/b - 1.75)$ (see e.g. Becker/Sauter, "*Theorie der Elektrizität*", Vol. 1, 21st ed., 1973, B.G. Teubner, Stuttgart); English: "Electromagnetic Fields and Interactions", Vol. I: Electromagnetic Theory and Relativity (Blaisdell, 1964). (Sect. 11.3)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_11) contains supplementary material, which is available to authorized users.

Electromagnetic Waves 12

12.1 Preliminary Remarks

In previous chapters, we organized the discussion of the electric field roughly as follows:

1. The *static electric field*. See the schematic in Fig. 12.1a. The field lines have *electric charges* at their ends.

2. A *slowly-varying electric field*. The two plates of a condenser are connected by a conducting bridge. In Fig. 12.1b, this is a long coil of wire. The electric field decays, but the self-inductance of the coil causes the decay to occur "slowly". The field decay occurs at α and β practically simultaneously. This is indicated in Fig. 12.1b by equal spacings of the field lines at α and β.

Now, in this chapter, we consider a third and final case:

3. A *rapidly-varying electric field*. In Fig. 12.1c, the bridge between the condenser plates is a short wire, so that its self-inductance is small. The field decays very *rapidly*: i.e. the time for the change in the field to propagate between α and β can no longer be neglected. The field decay caused by the bridge has already progressed much further at α than at β. That is indicated by the different spacings of the field lines in the figure. We thus observe a very high but still finite propagation velocity for electric fields. This finite propagation velocity makes it possible for *electromagnetic waves* to form. Such waves propagate either in all directions *freely*, like sound waves in open space, or else they are *conducted* along transmission lines, like the sound waves in a speaking tube.

Both forms of electromagnetic waves have yielded fundamentally important results for physics. First, their formation demonstrates experimentally that a time-varying electric field (the displacement current) also produces a magnetic field; this was made plausible by the MAXWELL equations, but was initially only a hypothesis (Sect. 6.4).

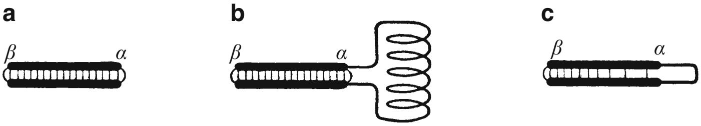


Figure 12.1 a: A static electric field; b and c: decaying fields within a condenser

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_12

Second, the discovery of electromagnetic waves has allowed us to extend the spectrum which was originally known only for visible light and infrared radiation seamlessly out to waves with wavelengths of up to many kilometers, and in the other direction to extremely short wavelengths in the regions of X-rays and gamma radiation.

In the realm of technology, both types of electromagnetic waves, free and conducted, have achieved an extraordinary importance. Modern communications technology, including television, satellite relays, mobile telephones, navigation methods for limited visibility conditions, etc. were all developed thanks to knowledge of electromagnetic waves. The methods devised for these purposes are increasingly permeating other areas of technology as well as our daily lives. The end of this evolution is not in sight. Only one thing is certain: All of these developments have grown out of physics and have become independent technological disciplines in their own right. Physics limits itself to the fundamentals. We should keep this in mind in reading the following sections.

12.2 A Simple Electrical Harmonic Oscillator Circuit

To demonstrate and investigate electromagnetic waves in the lecture room, we first need AC current sources with frequencies of around 100 MHz. These can best be generated using damped electrical oscillations. A suitable arrangement is shown in Fig. 12.2. Its disadvantage is obvious: The essential parts of the oscillator circuit, the condenser and the coil, are only vestigial, and they are practically invisible next to the trivial accessories which provide the feedback. We correct this in the well-known manner shown in Fig. 11.17: *Using an unspecified circuit ("black box") to provide the excitation, we produce resonant oscillations in a clear-cut setup*. The latter is in-

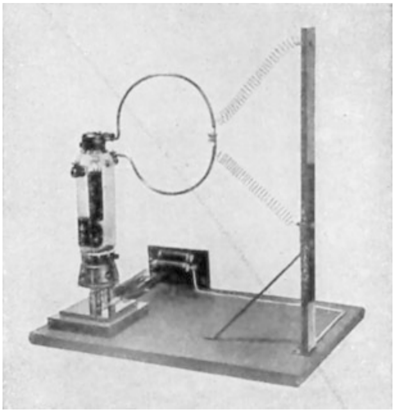


Figure 12.2 An oscillator circuit with feedback (frequency $\approx$ 100 MHz)

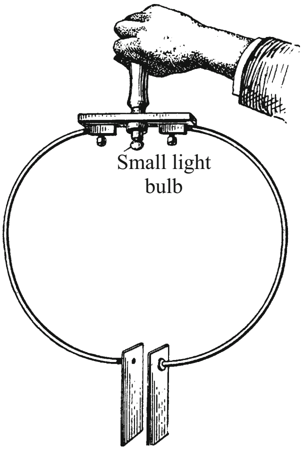


Figure 12.3 A simple closed-loop electrical oscillator circuit for demonstrating forced electrical oscillations. The light bulb serves as an indicator of the alternating current in the wire loop.

dicated in Fig. 12.3. We see just one simple circular copper wire loop of ca. 30 cm diameter. At its center, below the wooden handle, it contains a small light bulb as a current indicator. At each of its ends, a condenser plate about the size of a credit card is attached. The two plates are spaced around 5 cm from each other. This circuit serves as a *resonator*, and we bring it close to the excitation-source circuit shown in Fig. 12.2. By bending the copper-wire loop (and thus changing the spacing of the condenser plates), we can adjust the resonator frequency to be sufficiently close to the frequency of the excitation source. The light bulb glows brightly. In the resonator circuit, an AC current of about 0.5 A is flowing at a frequency of around 100 MHz.

Compare the experiments shown in Figs. 2.36 and 12.3. In Fig. 2.36, the field decays *once*, and the resulting current impulse is around 10^{-8} A s. In Fig. 12.3, the field decay occurs about 10^8 times per second, so that we can observe currents of the order of 1 A.

12.3 A Rod-Shaped Electric Dipole

With the high-frequency AC now at our disposal, we can proceed to something new and important, namely to a rod-shaped electric dipole or *antenna*.

In mechanics, a simple ball-and-spring pendulum consists of a body with inertial mass and an elastic component (helical springs, etc.). In electromagnetism, an electrical oscillator circuit with a coil and condenser is a correspondingly simple setup. We described the analogy of these two systems in Sect. 11.2, and refer here to Fig. 12.4.

In the simple ball-and-spring pendulum of mechanics, we can distinguish clearly between the inertial component and the component with

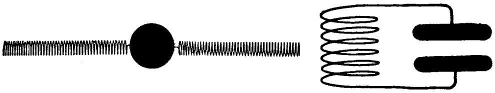


Figure 12.4 A simple mechanical ball-and-spring pendulum and the corresponding electrical oscillator circuit

elastic (restoring) force. As long as the mass of the ball is sufficiently large, we can completely neglect the small mass of the springs.

However, in mechanics we could also list numerous arrangements which can support oscillations or vibrations *without* a clear-cut separation and localization of the inertial component and the elastic component.[C12.1] An example is a column of air in a tube (an organ pipe). Every length element in the air column is both an inertial component and a section of 'tensed spring' (Vol. 1, Sect. 11.7).

C12.1. This kind of system, in which the components and their functions are spread over the whole system rather than being localized in specific devices, is called a "distributed system". In contrast, a system with separate, localized components and functions, like a ball-and-spring pendulum or a tank circuit with separate coil and condenser, is called a "lumped system".

A corresponding description applies to the case of electrical oscillations. In the typical tank circuit, e.g. as shown at the right in Fig. 12.4, we can clearly identify the coil as the location of the "inertia" of the magnetic field, and the condenser as the location of the "elastic" electric field. But in other electrical circuits which are capable of oscillating, localizing these functions separately is just as impossible as in the acoustically-vibrating air column in an organ pipe. An extreme case of this latter kind is a rod-shaped electric dipole. We will discuss this simple system in more detail.

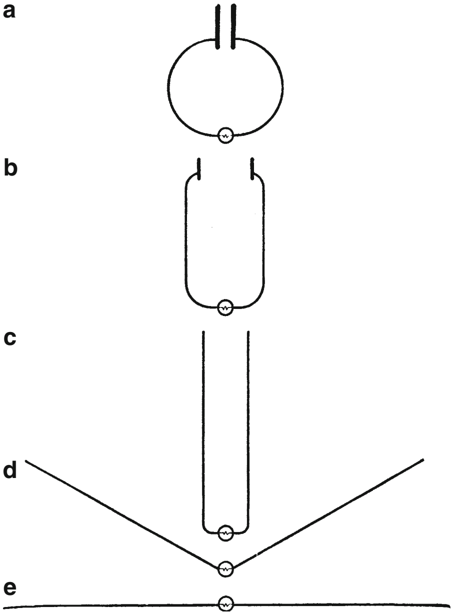


Figure 12.5 The transition from a closed ("lumped") oscillator circuit to an open, rod-shaped electric dipole ("distributed" circuit). The light bulb could be replaced by a suitable ammeter. It would indicate a current of around 0.5 A.

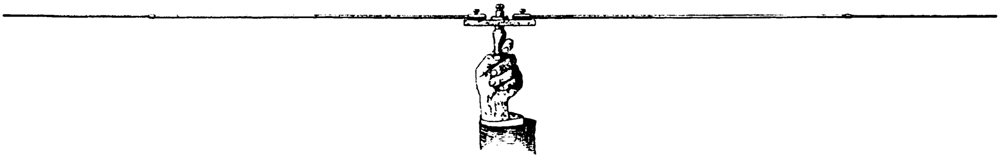


Figure 12.6 A rod-shaped electric dipole, around 1.5 m long

We again turn to the simplest of our oscillator circuits, the one shown in Fig. 12.3. When a current is passing through the wire loop and the light bulb, the electric field in the condenser is changing. We enlarge the region around the condenser, at the same time reducing the size of the condenser plates. We want to carry out the transition sketched in Fig. 12.5. The gradual atrophying of the condenser can be compensated by lengthening the two halves of the wire loop. The light bulb continues to glow, thus an AC current is still flowing.

In the limit, we arrive at Fig. 12.5e, a straight rod with a brightly glowing light bulb at its center. Fig. 12.6 shows how the experiment is carried out. The hand can serve as a length scale. The excitation source (Fig. 12.2) can be thought of as being about 0.5 m away.

The length of the rod is not very important. 10 cm more or less at each of its ends will play no significant role. The rod is a resonator with strong damping (Sect. 11.7). During its oscillation, the two halves of the rod are alternately positively and negatively charged. The corresponding charges can be thought of as each localized around a "center of gravity"; then we have two electric charge regions of opposite signs, separated by a distance l. An object of this kind was called an *electric dipole* in Sect. 3.9, and we adopt that name here for our electrically-oscillating rod. The electric charges oscillate back and forth along the rod. They represent conduction currents which alternate in direction, i.e. an alternating current. This is the electrical analog of an air current which alternates in direction within an organ pipe that is closed at both ends, a "stopped pipe": *In the dipole, electric charges are periodically accelerated, while in the organ pipe, air molecules are periodically accelerated.*

The fundamental oscillation of the pipe is shown in Fig. 12.7 by three "snapshot images". A grey shading refers to the normal number density of the air molecules, a darker shading is an increased density and a light shading a reduced density. These distributions are sketched graphically in the lower part of the figure. The wave (density or pressure) crests lie at the ends of the pipe and the node is in its center. These variations in the particle density occur because the individual volume elements of the air column in the pipe are flowing back and forth along it.

The air *currents* are also distributed sinusoidally, but their maximum (wave crest) is at the center of the pipe. There, the amplitudes of the velocities of the air molecules, which are directed alternately to the right and to the left, have their greatest magnitudes (Fig. 12.8).

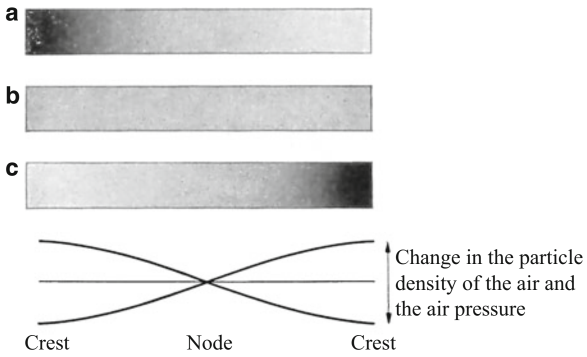

a
b
c
Change in the particle density of the air and the air pressure
Crest
Node
Crest

Figure 12.7 The distribution of the particle density of the air molecules and the air pressure in a pipe closed at both ends. Above, three "snapshots"; below, a graphical representation. Its ordinate, like the grey shading in the upper part of the figure, corresponds to the particle (number) density of the air molecules and at the same time to the air pressure.

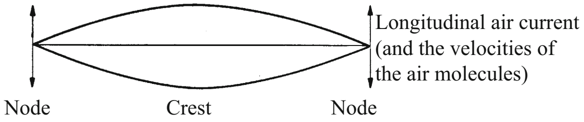

Longitudinal air current (and the velocities of the air molecules)
Node
Crest
Node

Figure 12.8 The sinusoidal distribution of the longitudinal air currents in a pipe which is closed at both ends

The electrical oscillations of a rod-shaped dipole behave in a corresponding manner. The air currents in the pipe are analogous to the electrical *conduction currents*[1], they are sinusoidally distributed along the length of the rod-shaped dipole. This is shown in Fig. 12.9 using three light bulbs: The middle bulb is glowing brightly, while the two on each side exhibit only a dull reddish-yellow glow. In Fig. 12.10, this sinusoidal distribution of the conduction current is shown as a graphical representation.

The periodically alternating current produces a periodically reversing electric-charge distribution. The grey shading in Fig. 12.7 corresponds to an electrically neutral state; the light shading is a positive, and the dark shading a negative charge excess. A positive charge excess makes the potential (the voltage between a point on the rod and for example the ground, or the center of the rod) positive, while a negative charge excess makes the potential negative. Figure 12.11 corresponds to the lower part of Fig. 12.7 for the organ pipe.

The analogy can be carried still further. The frequency of longitudinal mechanical vibrations is proportional to the square root of the modulus of elasticity E (Vol. 1, Sect. 11.8). For electrical oscilla-

[1] The electrons in these currents, due to their enormous numbers, move only through very short distances, on the order of a tenth of an atomic diameter.

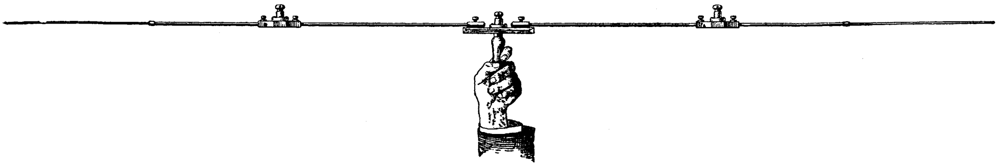


Figure 12.9 Demonstration of the sinusoidal distribution of the conduction currents along a rod-shaped dipole

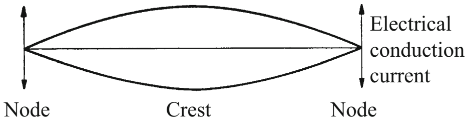


Figure 12.10 A graphical representation of the sinusoidal distribution of the conduction current along a rod-shaped dipole

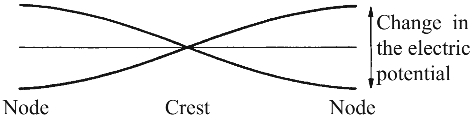
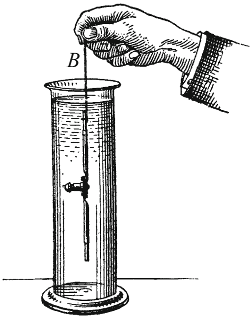


Figure 12.11 The distribution of the potential along the axis of a rod-shaped dipole. The abscissa line corresponds to a zero potential when the dipole as a whole is neutral.

tions, instead of the modulus of elasticity, we have the reciprocal of the capacitance C. The frequency of an electrical oscillation is proportional to $1/\sqrt{C}$. The capacitance C is itself proportional to the dielectric constant ε (Sect. 2.17). In a medium with a dielectric constant of ε, a dipole of length $l_\mathrm{m} = l/\sqrt{\varepsilon}$ has the same frequency as a dipole of length l in air or vacuum. This is shown in Fig. 12.12 for a dipole in water ($\varepsilon = 81$, $\sqrt{\varepsilon} = 9$, Table 13.3).

The correspondence between organ-pipe vibrations and dipole oscillations is still not exhausted. The pipe in Fig. 12.7 is vibrating at its fundamental frequency, while Fig. 12.13 corresponds to a pipe

Figure 12.12 In distilled water, this short dipole has the same resonance frequency as the dipole in air shown in Fig. 12.9, which is nine times longer (*B* is a piece of insulating twine) (**Exercise 12.4**)



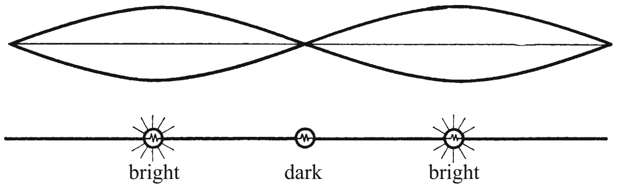


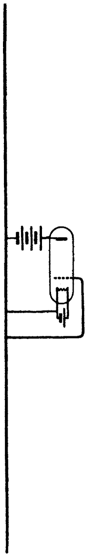
Figure 12.13 A dipole undergoing its first-harmonic oscillation, and its currents

which is vibrating at its *first overtone* (first harmonic, upper curve). In Fig. 12.13 below, a dipole about 3 m long is sketched; it was put together from two of the dipole segments used before. The light bulbs which have been inserted along its length allow us to visualize the distribution of the electric current: The light bulb at the central node remains dark, while the two on each side at the current maxima are bright. This dipole is also oscillating at its first overtone. In a corresponding manner, we could add sections to make dipoles with lengths of 4.5 m, 6 m, etc.

Just like an organ pipe in mechanics (acoustics), a dipole can of course also be excited to undamped oscillations by using autocontrol (feedback). This can be accomplished by using e.g. the circuit sketched in Fig. 12.14. It is derived directly from Fig. 11.13: The coil and condenser shown there have degenerated here into straight segments. The dipole with feedback has a refreshingly simple circuit diagram, but understanding it in detail unfortunately requires a knowledge of dipole oscillations.

So much for the rod-shaped dipole. It brought us an important bit of knowledge: The distribution of an *electrical conduction current*

Figure 12.14 A dipole with positive feedback (triode for autocontrol)



in the interior of a rod can take the form of a *standing wave*, in its fundamental as well as its harmonic oscillations.

This distribution of currents implies a particular distribution of the electric field. The investigation of this electric field and its time dependence is our next task. It will lead us to travelling electromagnetic waves, both those which are guided along a transmission line and those which propagate through free space.

12.4 Standing Waves Between Two Parallel Wires: The LECHER Line[2]

The electric field lines of an open, straight dipole must in some way be stretched in long arcs between various points along the length of the dipole. Along their paths, they meet up with the walls of the room, with the experimenter, etc. We will initially not attempt to describe this rather complex situation surrounding a straight, open dipole; instead, we begin by investigating the shape of the field lines in a simpler case.

In making the transition from a closed ("lumped") tank circuit to an open dipole, we passed through several intermediate stages, one of which is shown in Fig. 12.15 (top). We can call it a "folded" dipole for short. We bring it near to the excitation source at a frequency of 100 MHz (Fig. 12.2), and use the light bulbs to observe the distribution of currents in the dipole. The center lamp burns most brightly, and this is the location of the maximum current, the crest of the standing current wave.

With this dipole form, there can be no doubt about the shape of the electric field lines between the two "legs". The distribution of the electric field strength E is shown graphically in Fig. 12.15b. The two curves indicate the maximum values or amplitudes, similar to Fig. 12.11. In the upper curve, the upper half of the dipole has its maximum negative charge, while the lower curve shows the lower half with its maximum positive charge. The two curves follow each other at a time interval of one-half of a complete oscillation (cycle).

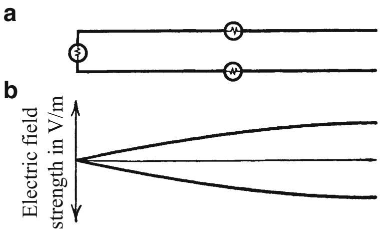


Figure 12.15 A "folded" dipole and the distribution of the electric field strength along its two legs

[2] E. Lecher, *Annalen der Physik* **41**, 850 (1890).

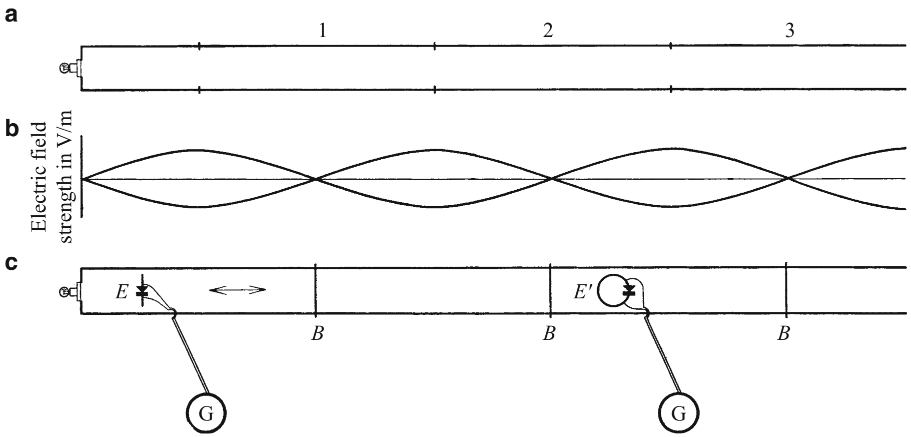


Figure 12.16 Demonstrating the standing electromagnetic waves between two parallel conductors (a Lecher line). When the field strengths are sufficiently high, small incandescent light bulbs can be used to sample the field distribution. The nodes of the magnetic field (which is perpendicular to the plane of the page) in the standing waves are at the positions of the crests in the electric field strength. Therefore, detectors for the magnetic fields must be well shielded against electric fields using a sheet-metal box (a Faraday cage).C12.2

C12.2. The standing wave of the magnetic field is thus shifted by 1/4 wavelength relative to the standing wave of the electric field! (Compare these field distributions with those of a travelling wave, Fig. 12.31.)

One can read the ordinates either as field strengths or as changes in the field strengths (displacement currents), since the regions of highest field strength are also regions where the field strength is changing most rapidly.

We could extend the dipole by adding one or more segments at its ends (Fig. 12.13). We see this in Fig. 12.16. The light bulb at the far left continues to burn brightly, i.e. the oscillations persist as before. The ends of the individual dipole segments are marked in Fig. 12.16a by short ticks. Below, in part b, the field distribution is again indicated; the field has its maximum strength at the crests and troughs, and drops to zero at the nodes (Fig. 12.16b).

This field distribution in such a system (called a Lecher system or Lecher line) can now be measured in an extremely simple and precise way. We mention two different procedures:

1. We can measure the strength of the electric field locally. A receiver (antenna in the form of a short dipole) can be located between the two legs; it is the short piece of wire labelled E in the figure. It “short circuits” the electric field, and an alternating current is produced within the wire by *influence*. We could also use an *induction* loop as a receiver (E'). It is permeated by the magnetic field which is directed perpendicular to the plane of the page and produced by the currents flowing in the line. The AC currents produced by influence (E), or by induction (E'), are converted to DC by a rectifier diode (detector) and can be measured with galvanometers (G). We move the receiver back and forth in the directions of the double arrow, and can thus locate the nodes, that is the zero points of the electric field, with consider-

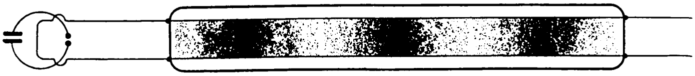


Figure 12.17 Visualization of the field distribution of standing electromagnetic waves between two parallel conductors (a LECHER line) (**Video 12.1**)

Video 12.1:
"LECHER line"
http://tiny.cc/gdggoy
In the video, the field distribution between the parallel lines is made visible in a simple way by using a fluorescent tube which is held adjacent to the LECHER line. The nodes of the electric field are always to the right of the insulating supports for the lines; the magnetic field has its oscillation maxima there.

able precision. Such methods are always applicable. We could then connect the two lines at the positions of the nodes with a finger or with a wire "bridge" *B* (compare Fig. 12.16c). This would not in the least perturb the standing waves: The light bulb at the left continues to glow with undiminished brightness.

We could "cut out" each of the rectangles bordered by neighboring bridges *B* and let them oscillate alone. To detect their oscillations, the bridges can be outfitted with small light bulbs. During the oscillation, periodic maxima of positive and negative charges are localized at the centers of the long, horizontal sides of the rectangles. Their charge transport back and forth through the two short sides (bridges) causes the light bulbs to glow.

2. Second method: We could pass the two parallel conductors through a long glass tube filled with neon at low pressure (Fig. 12.17). Then in the regions of high electric field strength (the crests and troughs of the standing waves), a gas discharge occurs in the neon; we can see the light from the positive column of the glow currents. The spatial alternation of dark and bright regions of gas in the tube gives a clear-cut image of the field distribution between the parallel conductors.

This method requires relatively high values of the electric field strength. They can be achieved most simply by using a damped excitation source, e.g. the circuit sketched in Fig. 12.17, with a spark gap (cf. Fig. 11.5).

The experiments described in this section lead to a simple but important result: *The electric field between parallel conductors can reproduce the form of a standing wave.*

12.5 Travelling Electromagnetic Waves Between Two Parallel Wires: Their Velocity

Standing waves are formed by superposition or interference of oppositely-directed travelling waves (Vol. 1, Sect. 12.5). Therefore, detection of standing electromagnetic waves proves that there are also *travelling* waves between the parallel conductors. A momentary image (snapshot) of such a wave is sketched in the upper part of Fig. 12.18.

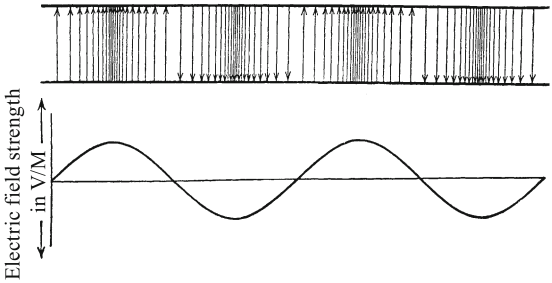


Figure 12.18 Top: A snapshot of a travelling electromagnetic "wire wave" between two parallel conductors (the arrows indicate the direction of the electric field, and their spacing shows the magnitude of the field strength). Bottom: A different representation of the snapshot of a travelling electromagnetic wave.

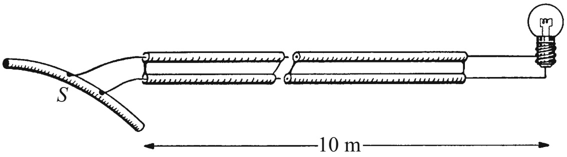


Figure 12.19 The conduction of a travelling electromagnetic wave ($\lambda = 3$ m) along a double line which is mounted on the edges of a plastic band 10 mm wide. The electromagnetic waves which are conducted along two parallel wires are called simply *alternating currents* when the length of the line is short compared to their wavelengths.

One can imagine that this entire image is moving in a horizontal direction with the velocity u. To an observer *at rest*, the travelling wave appears as a periodically-varying electric field.

At the bottom of Fig. 12.18 we see a different, but equivalent representation. Wave crests correspond to electric fields directed upwards, and wave troughs to electric fields directed downwards. The amplitude refers to the field strength in either case, quoted for example in V/m. But this graphical representation shows nothing of the shapes nor the extent of the field lines.

In Fig. 12.19, S is a segment of the oscillator circuit from Fig. 12.2. At two points, it is connected to a long double line (LECHER line) which has a light bulb at its far end. This lamp receives the energy from travelling electromagnetic waves that propagate along the line.

For all types of travelling waves, their frequencies ν, their wavelengths λ and their velocities of propagation u are related by the equation

$$u = \nu\lambda\,. \tag{12.1}$$

With a LECHER line, the frequency of the excitation system can be freely chosen. In principle, it can be calculated from Eq. (10.32). The wavelength λ can be measured; it is equal to twice the distance between two neighboring nodes. Inserting these values into Eq. (12.1) yields the velocity $u = 3 \cdot 10^8$ m/s = vacuum velocity of light, c.

This result is rather surprising: Along a LECHER line, every segment Δl, just as in any pair of conducting wires, has an OHMic resistance, an inductive resistance, and a capacitive resistance. Because of these resistances, the propagation velocity of sinusoidal waves along the line depends on their frequency. Non-periodic signals are thus collected together into wave groups (Vol. 1, Sect. 12.23). Their group velocity at frequencies of the order of several 100 Hz is found to be only around $2 \cdot 10^8$ m/s. Furthermore, the wave groups are damped along their paths. All of these technically important properties are described quantitatively by the so-called *telegraph equation*.[C12.3]

Why do all these properties of the transmission line play no role at the high frequencies of the LECHER system? Why do we find the full vacuum velocity of light, $c = 3 \cdot 10^8$ m/s, as the propagation velocity of the travelling waves in the limiting case of high frequencies? Answer: At high frequencies, the influence of the conduction currents in the conductors becomes completely unimportant. The magnetic field due to *displacement currents* – the large rate of change $\dot{E}$ of the electric field – is much stronger than the magnetic field from the conduction currents. It induces an electric field *between* the next segments of the line, and so forth.

The essential processes for the propagation of the waves thus do not take place within the conductors, but rather in the space between them; that is, in the air, or more strictly, in vacuum. Therefore, at high frequencies, the propagation velocity of the waves becomes independent of the properties of the conductors which make up the transmission line.

12.6 The Electric Field of a Dipole. The Emission of Free Electromagnetic Waves

Considering the results in the the preceding section, we can see that the parallel-conductor transmission line at high frequencies represents only an incidental accessory. It simply guides the waves so that they propagate along the line, rather than freely through space. It conducts the electromagnetic waves along a fixed direction just like a tube conducts sound waves in acoustics. Since this role is relatively unimportant, we can in the following forget about the transmission line. This will not hinder in any way the *essential* process, which is the mutual generation of alternating magnetic and electric fields due to their rapid time variations. We thus arrive at travelling electromagnetic waves which propagate freely in three-dimensional space. This brings us to our last and especially interesting question: How

C12.3. The telegraph equation is a second-order partial differential equation, whose solution describes the propagation of electromagnetic waves along a conducting transmission line. In addition to the inductance and capacitance of the line, it takes the OHMic resistance of the conductors into account. If the latter is negligible, the equation simplifies to the well-known *wave equation*, which can be derived directly from MAXWELL's equations. See for example B.E. Rebhan, "*Theoretische Physik*", Spektrum Akademischer Verlag, Heidelberg, Berlin 1999, Chap. 16; or R.P. Feynman *et al.*, "*Lectures on Physics*", Addison-Wesley, Reading, Massachusetts, 1964, Vol. II, Chap. 24 (available online; see Comment C6.1.)

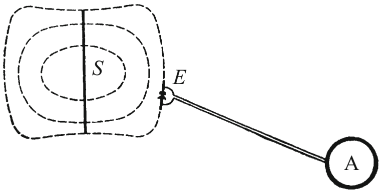


Figure 12.20 A snapshot of the electric field of a dipole S ("source"). A small dipole receiver (E) probes the field strength

are free electromagnetic waves emitted, how is their emission related to the accelerated motion of charges which are swinging back and forth (e.g. in an oscillating dipole)?

Our experimental starting point is once again the rod-shaped dipole. We briefly recall the distribution of conduction currents along the dipole. It exhibits a maximum at the center (Fig. 12.10). This current distribution corresponds to a particular distribution of the electric field. The field lines must somehow describe large arcs between corresponding points on the two halves of the dipole. Figure 12.20 shows a rough sketch for the case that there is maximum charge on the two ends of the dipole.

We now want to investigate the electric field of the dipole S in terms of its spatial distribution. We do this using the methods with which we are already familiar. We bring a short length of wire E to the position that we want to probe; a current will flow in the wire due to influence. We again call this probe the "receiver". The current in the receiver is alternating current with the frequency of oscillation of the dipole. A small rectifier diode (detector) converts the AC into a direct current, which we can readily measure using the ammeter A.

In order to avoid disturbances, we have to keep the distances to the walls of the room, the floor etc. large compared to the dimensions of the dipole. Thus we choose a dipole about 10 cm long. We employ damped oscillations for excitation, using a spark gap as switch (Fig. 11.3). The left part of Fig. 12.21 shows a convenient arrangement. The dipole consists of two identical thick brass rods. Their flat end surfaces are coated with magnesium foil.[C12.4] They are mounted with a gap of ca. 0.1 mm which forms the spark gap. A long, thin, flexible cable with two conductors (a household doorbell cable) provides the connection to an AC current source (around 5 kV, a small transformer operating at 50 Hz). At the terminals a and b, two small coils ("chokes") are included in the circuit; they isolate the supply cable from the high-frequency currents in the dipole. The spark gap makes no noticeable noise; we hear only a low humming. The dipole is mounted on a wooden column about one-half meter high. From now on, we will refer to it as the "transmitter" or *source* S. It can be tipped, rotated, or carried to another location at will during its operation.

The setup for the detection of the electric field remains the same as in Fig. 12.16. The receiver E has about the same length as the source S. This receiver is thus somewhat too large for measurements in the

[C12.4] The magnesium amplifies the sparks through which the alternating currents flow (Fig. 11.4).

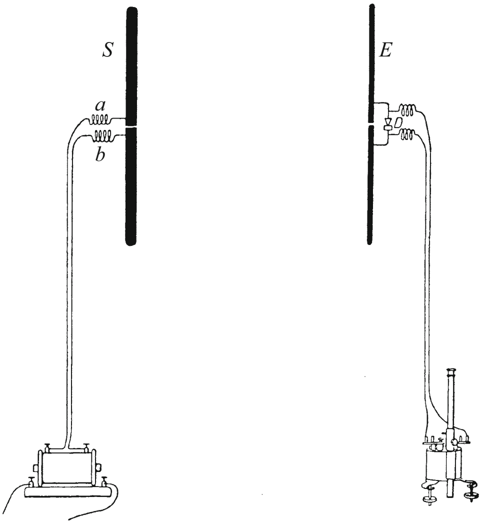


Figure 12.21 A small dipole as source S (at left) and as receiver E (right) (a and b are “choke’ coils, and D is a rectifier diode) (**Exercise 12.1**)

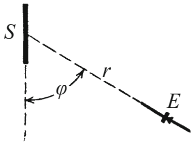


Figure 12.22 Measuring the radial component of the electric field (dipole field) in the neighborhood of the source dipole S

immediate neighborhood of the source; it would smear out the finer details of the field distribution there. This disadvantage of the relatively long receiver is compensated by its great sensitivity.

The receiver is also a dipole. It reacts to the oscillating field from the source by undergoing forced oscillations. The similarity of the lengths of the two dipoles guarantees that they can oscillate in resonance.

The receiver (Fig. 12.21, right) is connected to an ammeter by a thin, flexible pair of wires so that it is just as readily movable as the source. We can thus use it to sample the whole field distribution around the source.

We start by searching for *radial* components of the electric field *in the neighborhood of the source*; that is, we orient the source and the receiver as indicated in Fig. 12.22. These observations are carried out under various angles φ. In the neighborhood of the source, we find radially-directed electric field components at all values of the angle φ, but their magnitudes decrease rapidly with increasing distance r between the source and the receiver. At distances corresponding to two or three times the length of the dipoles, they are already negligible.

We continue our investigation, searching for *transverse components* of the electric field *close to the source*. We use the orientation illustrated in Fig. 12.23. These transverse components increase strongly with increasing values of the angle φ. But even for $\varphi = 0$, that is along the dipole axis of S, they still have noticeable magnitudes.

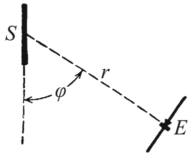


Figure 12.23 Measuring the transverse component of the dipole field

We then look for transverse components of the electric field at larger distances r from the source, five or six times the length of the dipoles. Here, we find no measurable transverse component along the direction of the dipole axis, i.e. at $\varphi = 0$. Such components can be found only at larger angles φ. For $\varphi = 90°$, the field strength is at a maximum; the field is transverse to the line r connecting the receiver to the center of the source dipole.

Thus far, we have kept the source and the receiver in the *same* plane, the plane of the page of Figs. 12.21 to 12.23. We now rotate either the source or the receiver slowly out of this plane: then the measured field strength decreases. It vanishes when the long axes of the source and the receiver are mutually perpendicular. The electric field $\boldsymbol{E}$ is a vector. According to our investigations, it lies in a plane together with the long axis of the source dipole.

At *greater distances*, the electric field exhibits a rather simple pattern, as seen from our observations; it can be represented graphically as illustrated in Fig. 12.24. The directions of the arrows indicate the direction of the electric field $\boldsymbol{E}$ at various sampling points which are at the same distance r from the dipole. The number of parallel arrows in the figure corresponds to the magnitude of the field, i.e. the field strength. The whole figure is just a small *section of a snapshot* of the electric field of the source dipole.

How would the *whole snapshot* image look? We can easily carry out the necessary extension of the image. We start by mentioning two facts:

1. The field pattern drawn in Fig. 12.24 comes exclusively from the source S (transmitter). It has propagated over the distance r through free space.

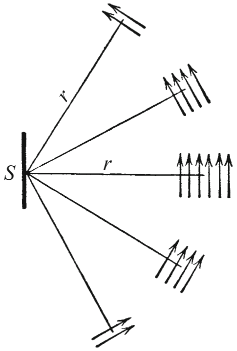


Figure 12.24 The distribution of transverse components of the electric dipole field in various directions

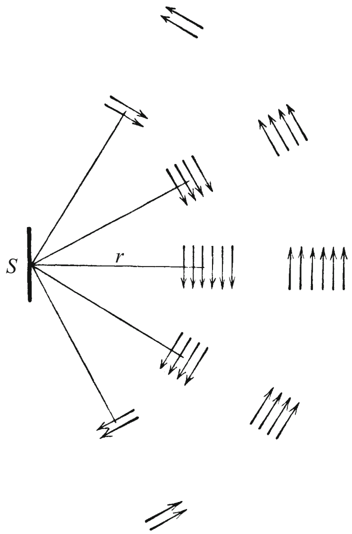


Figure 12.25 The spatial and temporal alternation of the electric field of an oscillating dipole

2. The field changes periodically at the frequency of the source. The snapshot in Fig. 12.24, if taken a short time later, would be replaced by a similar pattern but with the arrows reversed, that is with the *opposite direction* of the field. Such patterns alternate continually.

With these two facts, we can extrapolate the snapshot image in Fig. 12.24 as shown in Fig. 12.25.

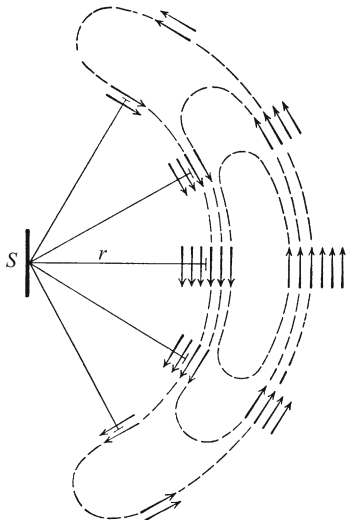
Now we take note of a third fundamental fact: Electric field lines cannot begin or end just anywhere in empty space. In empty space (without electric charges!), there can be only closed-loop electric field lines[3]. We have to extend the field lines as measured to give closed loops. This is shown in Fig. 12.26. We thus finally arrive at the complete snapshot image as in Fig. 12.27. It shows the electric field of the source dipole at some distance (i.e. excluding the near-field region). *This is the radiation field of the dipole, discovered by* HEINRICH HERTZ[4]. As a time-stopped image, it shows the emission of an electric field in the form of a transverse wave which is propagating outwards in free space. Its field strength is indicated by the density of the field lines as drawn. Imagine that the equatorial plane is drawn in and divided up into concentric rings of width $\lambda/2$. Then the surface density of the field lines decreases outwards in these rings as $1/r$ (r is the radius of the ring), *not* as $1/r^2$, like the radially-directed electric field of a charge at rest. *This is a fundamental difference between the electric field of an accelerated charge and that of a charge at rest.*

Figure 12.27, as we mentioned, represents a snapshot, a *time-stopped image*. Each radial segment of this image can be thought of as mov-

[3] The presence of the (relatively sparse) air molecules is quite unimportant for electric processes in space. We emphasize this point once again.

[4] *Annalen der Physik* **34**, 551, 610 (1888); *ibid.* **36**, 1 (1889).

Figure 12.26 The extension of the arrows in Fig. 12.25 to give complete, closed-loop electric field lines



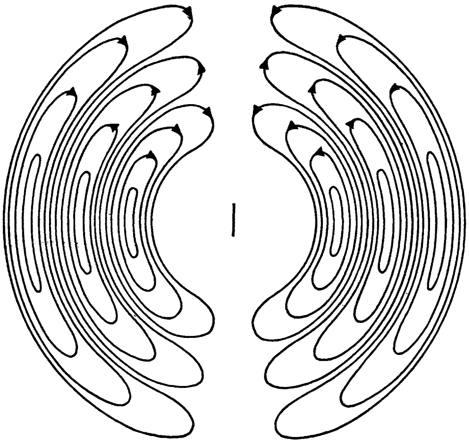
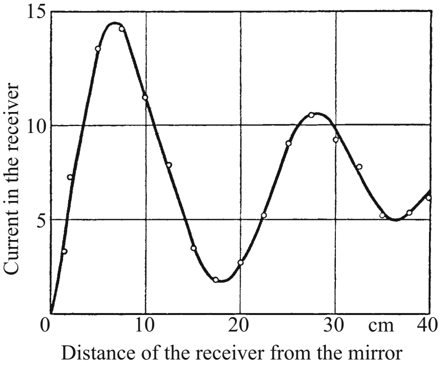


Figure 12.27 A snapshot (time-stopped) image of the electric field around an oscillating dipole: the HERTZian radiation field of a dipole. If we imagine this image to be rotated symmetrically around its vertical axis, the resulting three-dimensional distribution illustrates the $1/r$ dependence of the field strength (decreasing outwards).

ing outwards from the source at the velocity of light. This leads us to the vision of a wave which is propagating outwards from the dipole.

For the experimental detection of travelling waves, it is always useful to convert them into standing waves. We recall for example Fig. 12.44 in Vol. 1. Correspondingly, we allow the waves from the HERTZian dipole source to reflect at perpendicular incidence from a sheet-metal wall ("mirror") and move the receiver back and forth between the mirror and the source. We record the relative values of the current in the receiver using an ammeter, and thus the spatial

Figure 12.28 Measurement of the wavelength of the waves propagating outwards from the dipole shown in Fig. 12.21 (only relative values are given on the ordinate) (**Exercise 12.2**)



dependence of the field. The result of such a measurement is shown in Fig. 12.28. The nodes of the standing electromagnetic waves can be clearly seen as minima in the recorded field strength. The spacing of the nodes is found to be about 0.18 m. Therefore, the wavelength of the standing waves, and thus of the original travelling electromagnetic waves, is about 0.36 m in this example. From Eq. (12.1), the frequency ν of the dipole oscillations is given by:

$$\frac{3 \cdot 10^8\,\mathrm{m/s}}{0.36\,\mathrm{m}} \approx 800\,\mathrm{MHz}\,.$$

This experiment has a minor defect: The standing waves are clearly formed only in the neighborhood of the mirror. Further away, the minima become flatter and flatter. The reason for this is the strong damping of the oscillations of the source dipole. The individual wave trains initiated by the spark switch are too short; they resemble the curve shown for example in Fig. 11.18a at the upper right. At larger distances from the mirror, the strong displacements from the beginning of the wave train are superposed onto the weak displacements at the trailing end of the same wave train which are just moving out from the source. This yields only weakly discernable minima (cf. Optics, Sect. 20.11).

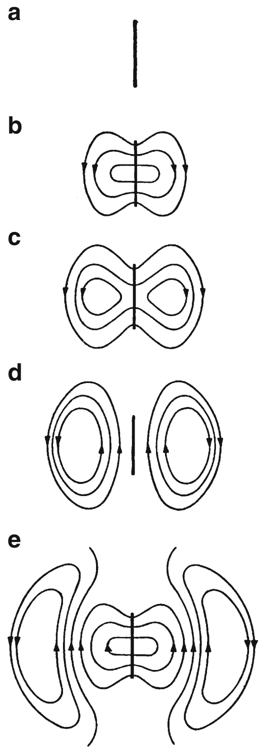
The pattern of wave emission from a dipole as sketched in Fig. 12.27 can thus be completely confirmed by experiment. An oscillating electric dipole emits travelling waves with their electric field perpendicular to their direction of propagation (transverse waves) into free space.

The field-line pattern of the dipole still requires two complementary additions:

In Fig. 12.27, the part of the field nearest to the dipole (the "near-field region") is not drawn in. In that region, it changes with the momentary charge state of the dipole. We give a brief description in the caption of Fig. 12.29.

Furthermore, the *magnetic field* of the dipole must be discussed. The magnetic field lines are concentric circles (Fig. 12.30). They lie in planes perpendicular to the dipole axis. The density and direction

Figure 12.29 Five snapshots of the electric field near an oscillating dipole: a. Before the start of the oscillations, both halves of the dipole are uncharged. There are thus no electric field lines connecting them. b. The conduction current has begun to flow upwards. After one-fourth of an oscillation cycle, it has charged the upper half of the dipole positive, and the lower half negative. Between the halves of the dipole there are now field lines which loop some distance outwards. c. During the second quarter-cycle, the magnitudes of the charges on both halves of the dipole again decrease: Here, they have already decreased by about half. The outer part of the field has moved further outwards, and at the same time, an odd constriction of the field lines near the dipole has appeared. d. At the end of the second quarter-cycle, the two halves of the dipole are again uncharged. The constriction of the field lines is complete. e. In the third quarter-cycle, the conduction current is now flowing downwards in the dipole, leading to a negative charge on its upper half and a positive charge on its lower half. At the end of the third quarter-cycle, the pattern resembles Part b, except that the arrows (field directions) are now reversed.



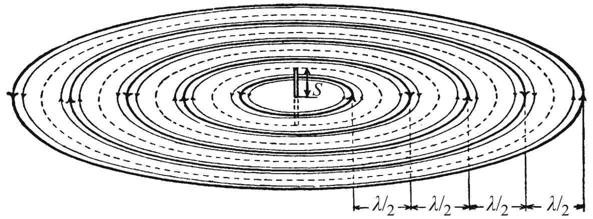


Figure 12.30 The magnetic field lines of an oscillating dipole

of the magnetic field lines alternate periodically. The magnetic field moves outwards together with the electric field; at a sufficient distance from the source ("far-field region"), it is in phase with the electric field.

Every variation in the electric field (displacement current) creates a magnetic field. And every variation of the magnetic field creates an electric field with closed-loop field lines, by induction. The propagation of the entire *electromagnetic wave* is based upon this close interlinking of the electric and the magnetic fields.

At a sufficiently great distance from the oscillating dipole, the wave can be described as a plane wave propagating in the direction of the

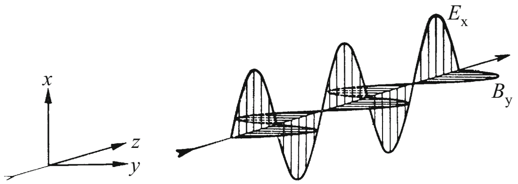


Figure 12.31 An electromagnetic plane wave propagating in the z direction consists of an electric and a magnetic component, which are polarized in the x and the y directions, respectively, and oscillate in phase$^{C12.5}$ (**Exercise 12.2**)

positive z axis at the velocity of light c (Fig. 12.31). Within this wave, the electric fields E_x oscillate in the x direction and the magnetic fields (or the magnetic flux density B_y) oscillate in the y direction. Expressed as equations, this corresponds to

$$E_x = E_{x,0} \sin\omega\left(t - \frac{z}{c}\right) \quad \text{and} \quad B_y = B_{y,0} \sin\omega\left(t - \frac{z}{c}\right) . \qquad (12.2)$$

The two waves thus oscillate in phase. The relation between their amplitudes is given by

$$B_{y,0} = \frac{E_{x,0}}{c} . \qquad (12.3)$$

Today, we can send free electromagnetic waves all around the globe. They propagate along great circles, experiencing multiple reflections on the upper layers of the atmosphere. These layers are ionized due to radiation from space, and thus have a high electrical conductivity (cf. Optics, Sect. 27.19). The earth's circumference of $4 \cdot 10^4$ km is traversed within 0.13 s, i.e. the waves could circle the globe seven times in 1 s. This makes possible a direct measurement of the propagation velocity of electromagnetic waves using their path travelled and their transit time.

C12.5. For completeness, we mention that the *energy* transported by an electromagnetic wave, i.e. the energy current density, is described by the POYNTING vector:

$$\boldsymbol{S} = \frac{1}{\mu_0}(\boldsymbol{E} \times \boldsymbol{B})$$

(unit: 1 W/m^2). With Eqns. (12.2) and (12.3), we find

$$S = \frac{1}{\mu_0 c} E_{x,0}^2 \sin^2 \omega\left(t - \frac{z}{c}\right)$$

which has the time-averaged value

$$\overline{S} = \frac{1}{2}\frac{1}{\mu_0 c} E_{x,0}^2 .$$

This quantity plays a role in optics.

12.7 Wave Resistance

For sound waves (Vol. 1, Sect. 12.24), we defined the quotient

$$\frac{\text{Pressure amplitude}}{\text{Velocity amplitude}} = \sqrt{\varrho K} = Z \qquad (12.4)$$

$$(K = \text{module of compression})$$

as the *acoustic wave resistance* for parallel beams of sound waves at normal incidence onto a boundary between two media. It determines the reflection coefficient at the boundary interface between the media. We found (Vol. 1, Eq. (12.50))

$$R = \frac{\text{Reflected radiation power}}{\text{Incident radiation power}} = \left(\frac{Z_1 - Z_2}{Z_1 + Z_2}\right)^2 \qquad (12.5)$$

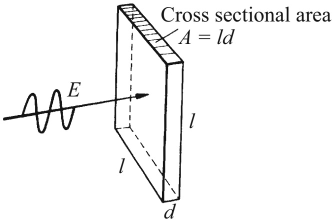


Figure 12.32 The calculation of the resistance $U/I = 1/\sigma d$ of a square piece of foil parallel to $\boldsymbol{E}$

for the *reflectivity* or the *reflection coefficient* for the sound waves.

In a completely analogous fashion, for electromagnetic waves we define the quotient

$$Z_{\mathrm{el}} = \frac{E}{H} \tag{12.6}$$

as the *wave resistance of vacuum* (Z_1 in Eq. (12.5)) for parallel beams of plane waves. For these waves, E and H are proportional to each other. From Eq. (12.3),

$$B = \frac{E}{c} = \sqrt{\varepsilon_0 \mu_0}\, E \quad \text{or} \quad H = \sqrt{\frac{\varepsilon_0}{\mu_0}}\, E\,, \tag{12.7}$$

it follows that

$$Z_{\mathrm{el}} = \sqrt{\frac{\mu_0}{\varepsilon_0}} = 377\,\Omega\,. \tag{12.8}$$

In Fig. 12.32, we show an electromagnetic wave which is approaching a poorly-conducting wall. We can imagine the wall to consist of a thin foil of thickness d made of some material with the specific electrical conductivity σ.C12.6 A square piece of such a foil is sketched in Fig. 12.32. Its OHMic resistance in the direction of the vector $\boldsymbol{E}$ is given by

C12.6. The specific conductivity σ is discussed in Comment C1.10.

$$\frac{U}{I} = \frac{1}{\sigma}\frac{l}{A} = \frac{1}{\sigma d}\,. \tag{12.9}$$

Foils are commercially available which as squares of arbitrary size have a resistance of $U/I = 1/\sigma d = 377\,\Omega$ (Z_2 in Eq. (12.5)). Such foils prevent the reflection of normally-incident electromagnetic plane waves.C12.7

C12.7. We mention a further application of Eq. (12.5): We calculate the reflectivity for an electromagnetic wave which is normally incident on the boundary surface between two dielectric media 1 and 2 (with the dielectric constants ε_1 and ε_2) (cf. Chap. 13). We have $Z_1 = \sqrt{\mu_0/\varepsilon_1\varepsilon_0}$ and $Z_2 = \sqrt{\mu_0/\varepsilon_2\varepsilon_0}$. Together with Eq. (12.11) from the following section, we find for the reflectivity

$$\left(\frac{Z_1 - Z_2}{Z_1 + Z_2}\right)^2 = \left(\tfrac{n_1 - n_2}{n_1 + n_2}\right)^2.$$

This result is derived later in Sect. (25.8) from FRESNEL's formulas.

In the case of *guided* or *conducted* electromagnetic waves, the value of the wave resistance depends on the shape of the electric field. In a LECHER system, for example, this shape depends on the diameter $2r$ and the spacing a of the two parallel conductors. Their wave resistance is

$$\frac{U}{I} = 120 \ln \frac{a}{r} \quad \Omega\,. \tag{12.10}$$

A numerical example: For $a/r = 16$, we find $U/I = 333\,\Omega$.

12.8 The Essential Similarity of Electromagnetic Waves and Light Waves

The HERTZian transmitter (Fig. 12.21, left) has a nearly ideal simplicity and clarity. With such a transmitter, we can readily demonstrate the analogous behavior of electromagnetic and optical waves. We have already observed reflection, interference and linear polarization. The vector of the electric field always oscillates within a plane which contains the axis of the transmitter dipole. A linear receiver perpendicular to this plane (Fig. 12.21 right) shows no signal.

HERTZ described a very impressive demonstration of this polarization of the dipole radiation, the so-called grid experiment[5]. We place the transmitter and the receiver parallel to each other. Then we insert a grid made of metal wires with a spacing of ca. 1 cm between the two. First, the wires are oriented perpendicular to the dipole axis and the direction of the electric field; this causes no noticeable weakening of the waves detected by the receiver. Then the grid is rotated by 90°, so that the wires are parallel to the dipole axis. It now proves to be completely opaque to the waves. The wires parallel to the direction of the electric field produce a "short circuit" and act just like a solid metal wall.
A similar experiment can be carried out in optics; however, for this we use invisible infrared waves ($\lambda = 100\,\mu\text{m}$). At the shorter wavelengths of visible light, it is not possible to fabricate a sufficiently fine wire grid; instead, one uses "polaroid foils" (Sect. 24.3). These contain oriented long-chain polymeric molecules with alternating double and single bonds, which are electrically conducting along their long axes and thus act as nanoscopic "wires".

For the demonstration of refraction, a "cylindrical lens" suffices. This is a large bottle which is filled with a dielectric liquid, e.g. xylol. Its long axis is oriented parallel to the receiver dipole. Using prisms of adequate size, HERTZ was able to determine the refractive index n of several substances in his classical experiments with electromagnetic waves. He found that n is equal to the square root of the dielectric constant ε of the substance of which the prism is made. This relation, $n = \sqrt{\varepsilon}$, had already been predicted by MAXWELL on the basis of his electromagnetic equations. It plays an important role in the theory of dispersion (Optics, Sect. 27.8).

MAXWELL's relation $n = \sqrt{\varepsilon}$ follows from Eq. (8.7). In a material with the dielectric constant ε and the permeability μ (this will be treated in Sect. 14.1), the products $\varepsilon\varepsilon_0$ and $\mu\mu_0$ replace ε_0 and μ_0 in MAXWELL's equations. We thus obtain

$$\text{Index of refraction } n = \frac{c_{\text{vacuum}}}{c_{\text{material}}} = \frac{\sqrt{\varepsilon\varepsilon_0\mu\mu_0}}{\sqrt{\varepsilon_0\mu_0}} = \sqrt{\varepsilon\mu}\,. \qquad (12.11)$$

The permeability of most materials, apart from ferromagnets, is practically always $\mu \approx 1$ (Sect. 14.3); we thus obtain $n = \sqrt{\varepsilon}$ (**Exercise 12.4**).

[5] *Annalen der Physik* **36**, 769 (1888).

12.9 The Technical Importance of Electromagnetic Waves

The technology of electronic communications, from its earliest beginnings in the first half of the 19th century, for many decades used direct currents as *carriers*. They were modulated by employing switching devices, e.g. a telegraph key, or by utilizing microphones. The transmission lines then carried a chopped direct current or an alternating current in the frequency range of human speech (audio frequencies, audio currents).

Beginning in 1896 (G. MARCONI), modulated electromagnetic waves have been increasingly used as communications carriers. Initially this was achieved with freely propagating waves travelling in all directions (Fig. 2.12 shows schematically an antenna designed to receive such waves). Later, the waves were *focussed into beams* using concave mirrors (as in a searchlight) or were *guided* along double-conductor lines (a further development of the LECHER line). In the course of these developments, shorter and shorter wavelengths were employed. For their great and often admirable achievements – think of the transmission of images of the surfaces of Mars, Jupiter etc.! – communications engineers had to develop new technologies for sources of undamped electromagnetic waves with wavelengths down to centimeters (today, down to millimeters) and for their propagation and their precise control. These technologies have also enriched physics laboratories for research and teaching. The next sections will offer some examples of these developments as applied to fundamental physics.

12.10 The Production of Undamped Microwaves. Demonstration Experiments for Wave Optics

For generating undamped electrical oscillations using normal electron tubes or triodes, an electron current flows in the tube between its cathode and the anode (plate). The charge (number) density of the space charge is periodically varied (*modulated*) using a control voltage applied between the grid and the cathode. In this way, even with special designs, it is not possible to generate oscillation frequencies above $3\,\text{GHz} = 3 \cdot 10^9\,\text{Hz}$, owing to the finite travel time of the electrons which make up the space charge. This means that the lower limit for the wavelengths of the electromagnetic waves generated in such tubes is around 10 cm. However, this travel time itself can be employed to achieve frequencies up to 100 GHz, i.e. wavelengths in the mm range. We describe here one of the several methods which have proved to be practical, based on the *klystron*.

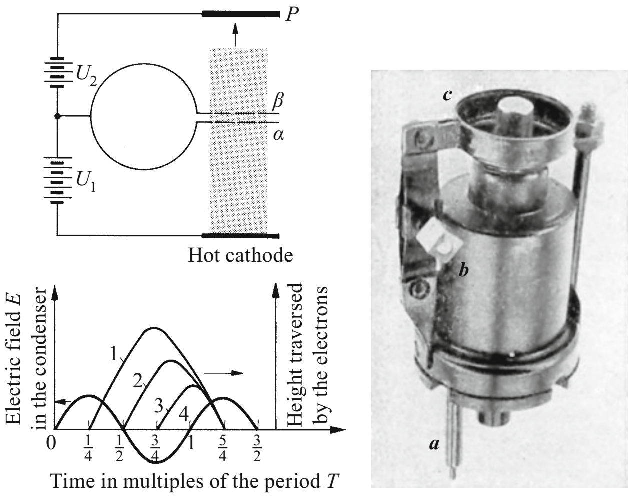


Figure 12.33 The autocontrol (feedback) of electrical oscillations using a reflection klystron. At the left, a schematic; at the right, a photograph of a technical device, with rotational symmetry, for frequencies of around 10 GHz, which can be regulated by varying the distance $\beta - \alpha$ (screw at b) (ca. 2/3 actual size; the operating voltages are $U_1 \approx 300$ V, $U_2 \approx 150$ V; a indicates a coaxial cable connector for outputting the electromagnetic waves).

Figure 12.33 (upper left) shows an oscillator circuit schematically. The plates of its very flat condenser (at α and β) are perforated like sieves. Electrons which are emitted by a hot cathode (below) can pass through the openings. Their flight times t within the gap of the condenser are short compared to the period T of the oscillator circuit. Furthermore, suppose that some quite weak oscillations are already present due to random fluctuations (from thermal motions). Then bunches of electrons leave the condenser at its top with a *modulated velocity*. Within each period T, when the upper condenser plate is *positive*, the emerging electrons have higher velocities, up to $(u + \mathrm{d}u)$ at the time $T/4$. The electrons which emerge later, at the time $T/2$, when the upper plate of the condenser is *not charged*, have the same velocities u with which they entered through the lower plate. Still later, when the upper condenser plate is *negative*, the emerging electrons have reduced velocities, down to $(u - \mathrm{d}u)$ at the time $3T/4$.

During this modulation of their velocities, the *number density* of the electrons remains constant. In order to modulate it as well, we let the electron bunches be reflected by a negatively-charged plate P, so that they return to the condenser. The fastest electrons (with velocities up to $(u + \mathrm{d}u)$) have to traverse the longest paths upwards and downwards; electrons with the original velocity u traverse a path of medium length, and the slowest electrons (with velocities down to $(u - \mathrm{d}u)$) traverse the shortest paths (Fig. 12.33, lower left).

The fastest electrons of the first group begin their flight upwards at the moment when the positive charge on the upper condenser plate is at its maximum, i.e. each time at $T/4$; the electrons of the second group, with the velocity u, start each time at $T/2$, when the condenser plate is uncharged; and the slowest electrons of the third group start each time at $3T/4$, when the upper condenser plate has its maximum negative charge.

If the voltages U_1 and U_2 (Fig. 12.33, upper left) are properly chosen, then two conditions are met. First: After each full cycle (beginning each time at $T/4$ and ending at $5T/4$), all the electrons arrive back at the upper condenser plate nearly *simultaneously*. Second: They pass downwards through the condenser in *packets with their modulated number densities combined*. This occurs periodically, always at times when the *lower* condenser plate is negative. As a result, the electron packets are slowed; they give up a portion of their kinetic energy to the electric field of the condenser. This periodic energy input initially amplifies the weak oscillations, and then maintains them at a constant amplitude.

Figure 12.33 (right side) shows a wave generator based on this method, for wavelengths of ca. 3 cm. a is a coaxial connector where the high-frequency alternating current ($\nu = 10\,\text{GHz}$) can be coupled out of the generator, e.g. for feeding a transmitter antenna. With this transmitter, and a receiver antenna with a detector and ammeter, we can demonstrate all the basic phenomena of wave propagation, at a somewhat higher cost but no less conveniently than with short-wavelength sound waves (Vol. 1, Sects. 12.18 through 12.20). We add to the examples given in Vol. 1 by demonstrating a lens in Fig. 12.34. Speaking figuratively, it is made as a regular crystal. Its "atoms" consist of thumbtacks ordered on a square lattice. More details are given in the figure caption.

Among the other devices used to generate undamped electrical oscillations at very high frequencies which are based on the finite transit times of electrons, the most important is the *magnetron*. There, the

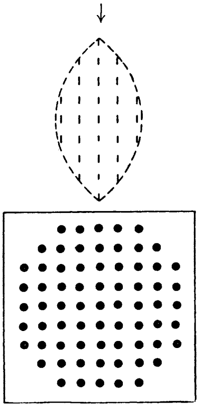


Figure 12.34 *Above*: A cross section through a lens for electromagnetic waves. The scattering elements or "atoms" are thumbtacks which are pressed into a polystyrene plate (transparent to the waves). *Below*: The plate in a plan view (the section marked above with an arrow). (See Sect. 27.6)

energy required to maintain the high-frequency oscillations at constant amplitude is provided by electrons that are moving on circular orbits at their cyclotron frequencies in a magnetic field (see Comment C8.1). With such magnetrons, which played an important role in the early development of radar, electrical oscillations at frequencies between 1 and 100 GHz can be generated, corresponding to wavelengths between 30 cm and 3 mm. Magnetrons with an output power in the range of several kilowatt are used today in microwave ovens (see Sect. 13.11).

12.11 Waveguides for Short-Wavelength Electromagnetic Waves (Microwaves)

In a LECHER line, the distance between the two conductors is *small* compared to the wavelength of the electromagnetic waves they conduct. The LECHER line thus corresponds to the *speaking tubes* which were once common in households and in ships. The LECHER system can also be redesigned: One of the conductors can be formed as a tube which surrounds the other concentrically, and the two conductors can be separated by dielectric spacers or a dielectric material which fills all the space between the outer conductor (tube) and the inner central wire.[C12.8]

C12.8. However, when the space between the conductors in such a *coaxial cable* is filled with a dielectric material (plastic) with an index of refraction n, the waves propagate along the cable only at the velocity of light in that medium, that is at c/n.

Starting with such a concentric, coaxial LECHER system, one arrives at a *waveguide* when the axial (center) conductor is left out completely (LORD RAYLEIGH, 1897)[6]. Then only the outer tube remains. In practice, usually a rectangular cross section is employed. A waveguide has quite different properties from a speaking tube or its electrical analog, a LECHER line: A waveguide transmits only those waves whose half-wavelength is *smaller* than the largest inner dimension of the waveguide. Here, we must distinguish between two different velocities: First, the velocity u at which a signal carried by the waves, that is a wave train or packet with a beginning and an end, called a *wave group*, can move on a zig-zag path along the axis of the waveguide. Second, the phase velocity v at which a wave which is modulated transversely to the axis travels through the waveguide. These two velocities are related to the velocity c of electromagnetic

[6] Concentric LECHER lines have the disadvantage that they often function *in addition* as waveguides. This can be prevented only by reducing the cross-section of the outer tube. This reduction, however, leads not only to technical difficulties (centering of the axial conductor, sufficient electrical insulation), but also to an unacceptably large increase in the damping. The latter increases when the quotient circumference/cross-sectional area increases.

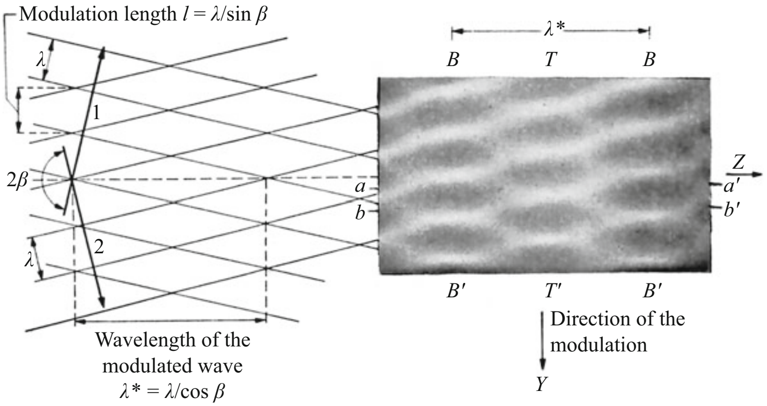

Modulation length $l = \lambda/\sin\beta$
λ
1
2β
2
λ
Wavelength of the modulated wave
$\lambda^* = \lambda/\cos\beta$
λ^*
B
T
B
a
b
Z
a'
b'
B'
T'
B'
Direction of the modulation
Y

Figure 12.35 Please look at this figure with one eye! The interference between two plane waves of the same frequency which meet at an angle 2β produces travelling waves in the z direction. Their amplitudes, that is the heights of their crests and troughs, are directed along the y axis, and are thus *transverse* to the propagation direction; they are modulated with the modulation length $l = \lambda/\sin\beta$ (at the right is a snapshot of waves on a water surface, taken with an exposure time of (1/250) s). As the angle β between the two wave trains increases, so does the phase velocity $v = c/\cos\beta$ of the resultant waves propagating along the z direction, whose amplitudes are transversely modulated. This can be readily seen in the demonstration experiment. At the limiting angle $\beta = 90°$, we find for the phase velocity $v = \infty$; the result is a *standing* wave. Its modulation length is $l = \lambda$; that is, the spacing between neighboring interference minima is $l/2 = \lambda/2$. In standing waves, the interference minima are called *nodes*. (In the example: $\beta = 76°$, $\lambda = 5.8$ mm, $l = 6$ mm, $\lambda^* = 25$ mm)

waves in free space (vacuum)[7] by the equation

$$uv = c^2 . \tag{12.12}$$

These properties of waveguides, which at first may appear somewhat strange, have a single origin: The electric field strength must be zero wherever the electric field lines of the wave approach the metal walls. We must explain this in somewhat more detail:

As an introduction, Fig. 12.35 shows a well-known phenomenon from mechanics. At the left is a sketch of the *momentary* configuration, and at the right a photographic image, of surface waves on water resulting from interference between two linear wave trains. Their directions of propagation 1 and 2 make an angle of 2β with each other. The two wave trains superpose to give a resultant wave train; it has the wavelength $\lambda^* = \lambda/\cos\beta$ and travels in the z direction at the high phase velocity $v = c/\cos\beta$. Its amplitude, that is the height of its wave crests (e.g. along BB') and the depth of its troughs (e.g. TT'), vary along the y direction, i.e. the wave is thus *modu-*

[7] Free space (vacuum) is non-dispersive, and therefore, its signal or group velocity and its phase velocity are identical.

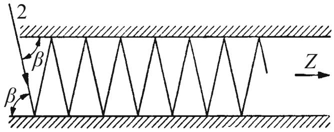


Figure 12.36 Construction of a wave which propagates in the z direction and is modulated *transversely* to its direction of propagation. It utilizes the multiple reflections of a plane wave on two perfectly reflecting walls placed at the interference minima. β has the same meaning here as in Fig. 12.35. Along the zig-zag path, energy is transported at the velocity c, while along the waveguide axis (z direction), it is transported only at the group velocity $u = c\ \cos\beta$. The figure also explains the propagation of energy in any interference field: The interference minima act, figuratively speaking, as perfect mirrors.

lated transversely[8] to the propagation direction: The wave crests are divided by depressions and the troughs by elevations which follow each other at a spacing $l = \lambda/\sin\beta$ (called the *modulation length*). *Half* of this modulation length is the distance between neighboring *interference minima*, i.e. the lines along which the amplitude remains zero (for example the lines connecting the points aa', bb' etc.).

The wave sequence discussed in Fig. 12.35 can be produced experimentally in a simple fashion. This is shown in Fig. 12.36: The region where the waves can propagate freely is bounded above and below by two perfectly reflecting walls (shaded). The spacing of these walls is taken to be an integral multiple N of half the modulation length, i.e. $l/2$. At the left, a *single* wave train enters the region between the two mirrors, and it passes along them at the phase velocity c (see footnote 2), following a zig-zag path. As a result, a signal propagates in the z direction only at the low velocity $u = c\ \cos\beta$. We thus find, since the phase velocity of the resultant wave with wavelength λ^* is $v = c/\cos\beta$, that $uv = c^2$.

So much for the results with mechanical (surface) waves. Their application to electromagnetic waves in waveguides need only be elucidated with the help of an example. Figure 12.37 shows a first approximation: Instead of a rectangular waveguide, we see a region in the field of a linearly-polarized electromagnetic plane wave which is bounded by two conducting sheets. The wave propagates in the z direction with its electric field in the x direction. The amplitudes of the wave are independent of the y direction, apart from some unimportant edge effects; therefore, we can represent the amplitude all along the y axis with arrows of the same length. This is shown in the figure at two amplitudes.

Figure 12.38 shows a different case: Here, the two vertical conducting sheets are joined above and below by horizontal sheets to make a closed profile with a rectangular cross section. Now, a homogeneous distribution of the electric field in the y direction is impossible: The electric field has to be zero above and below where its field lines

[8] In communications technology, an amplitude modulation of waves *along* their propagation direction also plays a significant role.

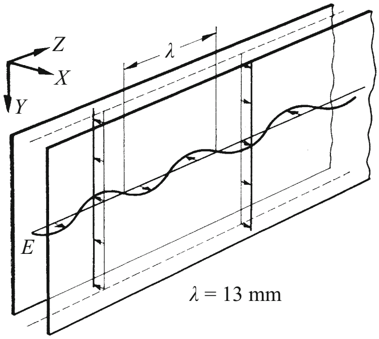


Figure 12.37 A drawing in perspective of a region bounded by two conducting sheets in which an electromagnetic plane wave is propagating with its electric field oscillating in the *x* direction. Its field lines meet the two sheets at normal incidence and end there. The electric field amplitude is independent of *y*.

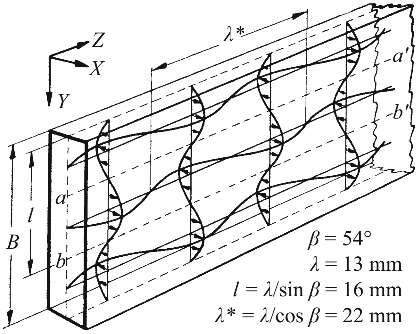


Figure 12.38 The corresponding picture after the two vertical conducting sheets have been joined above and below by horizontal sheets to give a rectangular profile, a RAYLEIGH waveguide. The waves must be modulated in the *y* direction, that is perpendicular to their direction of propagation, so that their electric field strength becomes zero at the points where the field lines intersect the conducting walls.

graze the conducting walls (corresponding to a perfect reflection). This is achieved by a *modulation* of the wave amplitude in the *y* direction. In Fig. 12.38, the modulation length has been chosen to be $l = 2B/3$ (corresponding in the example to $\beta = 54°$). In the case of the water surface waves, Fig. 12.38 would correspond to Fig. 12.39. Waves can propagate through this waveguide only when a modulation length of $l = 2B/N$ can be achieved for them within the waveguide (N is an integer). Therefore, $\lambda = 2B$ is the largest wavelength which can be transmitted through the waveguide. Longer waves cannot fulfill this requirement, namely that the electric field be exactly zero at the points where its field lines graze the conducting walls of the waveguide.

Instead of a tube with a circular or rectangular cross section, a number of other forms can be used as waveguides. For example, a helical wire, a single wire with a dielectric covering, or in the laboratory and for demonstration experiments, a rubber hose[9] might be used.

[9] F.E. Borgnis and C.H. Papas, "Electromagnetic Waveguides", *Handbuch der Physik*, Flügge, Vol. 16 (1958).

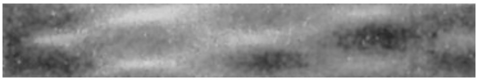


Figure 12.39 The same field distribution as in Fig. 12.38, here as a section from a photograph of water surface waves. The saddle points along the wave crests correspond to regions with field strengths $\boldsymbol{E}$ pointing in the positive x direction, while in the troughs, $\boldsymbol{E}$ points in the negative x direction. At rest, the water surface would correspond to the area bounded by dashed lines in Fig. 12.38, where the vector arrows representing the electric field all begin.

For measurements and experiments which use electromagnetic waves in the centimeter and decimeter wavelength range, waveguides are just as important as wires and coaxial cables for the usual electrical measurement instrumentation. There is a voluminous specialized literature with its own terminology. For example: Reflections on the tube walls are decisive for the formation of tube waves. As is known from optics, reflections are dependent on how the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{H}$ are oriented with respect to the plane of incidence (Optics, Sect. 25.6). If $\boldsymbol{E}$ lies perpendicular to the plane of incidence[10], i.e. it is "transverse", then $\boldsymbol{H}$ has a component parallel to the long axis of the tube. If $\boldsymbol{H}$ lies perpendicular to the plane of incidence ("transverse"), then $\boldsymbol{E}$ has a component parallel to the tube axis. In the former case, waveguide engineering classes the waves as TE or M waves; in the latter case, as TM or E waves. Often, indices are also added to denote the number of modulation lengths parallel to the long or short sides of the rectangular waveguide.

For measurement purposes, a narrow *longitudinal slit* is cut into the wall of the waveguide. Then a small receiver (cf. Fig. 12.16) can be slipped in, parallel to the waveguide axis. It can be used to determine the wavelength when a reflecting wall at the end of the waveguide is used to produce standing waves. Thus, for example, we can measure e.g. the wavelength λ^* in Fig. 12.38 which belongs to the high phase velocity $v = c/\cos\beta$.

Exercises

12.1 In Fig. 12.21, the ammeter indicates a current, although of course it responds only to currents of very low frequency. Explain this. (Sect. 12.6)

12.2 In order to understand the observed standing electrical wave in Fig. 12.28 as the superposition of two travelling waves, consider an electromagnetic wave which is incident along the z axis from $z > 0$ onto a metal mirror at $z = 0$. The motion of its electric-field component is described by $E_{\mathrm{x}} = E_{\mathrm{x},0} \sin\omega(t + z/c)$. At the mirror, this wave

[10] The plane of the page in Fig. 12.36, the yz plane in Fig. 12.38.

is reflected and it propagates in the positive z direction, opposite to the incident wave. The two wave trains superpose to give a resultant wave E_{tot}. a) Write the equation which describes the reflected wave, so that, as in Fig. 12.28, a node is present at the mirror ($z = 0$). b) In addition to the electric-field component E_{x}, electromagnetic waves have also a magnetic component B_{y}. It oscillates in phase with the electric component (Eq. (12.2)), where E, B and the propagation direction are oriented to form a right-handed coordinate system (see Fig. 12.31). The reflection leads to a standing wave B_{tot} for the magnetic component, also. Find expressions for the two magnetic-field waves which are superposed, and for B_{tot}. What is the amplitude of B_{tot} at the surface of the mirror? (The trigonometric formula for the sum of two sine functions is helpful for solving this exercise.) (Sect. 12.6)

12.3 The heating element of a hotplate is 1 m long and has a diameter of $2r = 1\,\text{cm}$. a) At a voltage of 220 V and a current of 4.5 A (effective values), calculate the electric field E in the element, the flux density B of the magnetic field at its surface, and the resulting POYNTING vector $\boldsymbol{S}$ (see Comment C12.5.)
b) Compare the values obtained with that of the sun's radiation (solar constant b, Sect. 19.3). (Sect. 12.6)

12.4 In Fig. 12.12, a dipole in water is excited to electrical oscillations by electromagnetic waves which are radiated into the glass vessel through its surface at normal incidence from the outside. Find the reflectivity R of the water in this experiment (see Comment C12.7). (Sect. 12.8)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_12) contains supplementary material, which is available to authorized users.

Matter in an Electric Field 13

13.1 Introduction. The Dielectric Constant ε

Thus far, we have dealt only with electric fields in empty space. The presence of air molecules could be neglected in our considerations; their influence is felt only in the fourth decimal place by 6 digits. However, when we consider fields in insulating materials – *dielectrics* – with a dense packing of atoms or molecules, that is liquids and solids, the situation becomes different. Such a dielectric between the plates of a condenser increases its capacitance. Thus, in Sect. 2.17, we proposed a definition of the *dielectric constant*:C13.1

$$\varepsilon = \frac{\text{Capacitance } C_\text{m} \text{ of a condenser which is filled with matter}}{\text{Capacitance } C_0 \text{ of the empty condenser}} \,. \tag{2.25}$$

C_m is larger than C_0, so that ε is always a number greater than one.

C13.1. In general, ε is not constant, as is demonstrated in the following with a number of examples. In the literature, ε is often denoted as ε_r (r means "relative to the vacuum") and is called the "relative permittivity". The product $\varepsilon_0\varepsilon_\text{r}$ is frequently denoted simply by ε and is called the "permittivity" (and therefore, ε_0 is the "permittivity of vacuum").

13.2 Measuring the Dielectric Constant ε

At a fixed voltage U, the charge Q on a condenser is proportional to its capacitance C ($C = Q/U$). For a parallel-plate condenser, the quotient of the charge Q and the plate area A gives the surface charge density or displacement density $D = Q/A$ (Sect. 2.13). Then we obtain for the dielectric constant:

$$\varepsilon = \frac{Q_\text{m}}{Q_0} = \frac{D_\text{m}}{D_0} \,. \tag{13.1}$$

The charge, and its surface density, increase when a dielectric is inserted. Figure 13.1 shows how we can measure ε based on this fact.

Various experimental arrangements have been developed for the measurement of the dielectric constant. Usually, instead of only one current impulse or charging cycle of the condenser, a periodic series of current impulses is employed; this is achieved by utilizing alternating current. Furthermore, the sensitivity of the measurement is

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_13

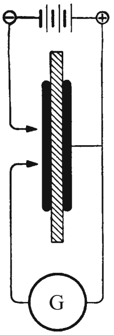


Figure 13.1 Measuring the dielectric constant ε. The increase in the charge on the condenser plates is determined as a current impulse by a ballistic galvanometer. The deflection of the galvanometer is proportional to ε.

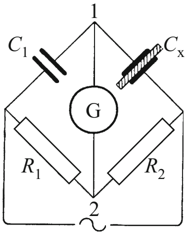


Figure 13.2 A bridge circuit for comparing two capacitances. The measuring instrument (e.g. a galvanometer or an oscilloscope) is used as a zero detector. The OHMic resistors R_1 and R_2 are variable and known; the capacitance C_1 is also known (**Exercise 13.3**)

improved by using some sort of *compensation* or *difference* method, for example some type of "bridge circuit" (Fig. 13.2).

Equation (13.1) holds for a fixed voltage. When the voltage source is disconnected, the charge Q remains constant (Fig. 2.3). Then C is proportional to $1/U$. For a parallel-plate condenser, the quotient of voltage U divided by the plate spacing l gives the electric field strength, $E = U/l$ (Sect. 2.12). Then we have for the dielectric constant:

$$\varepsilon = \frac{U_0}{U_\mathrm{m}} = \frac{E_0}{E_\mathrm{m}} . \tag{13.2}$$

The voltage or the electric field strength thus decreases when a dielectric is inserted; for the measurement, we determine the voltage drop when the condenser is completely filled with the material. It is proportional to ε (**Video 2.1**)

Video 2.1:
"Matter in an electric field"
http://tiny.cc/s9fgoy.
It is shown how inserting various materials into the electric field of a parallel-plate condenser which has been disconnected from the current source produces a decrease in the condenser voltage. Its capacitance is thus increased. In this experiment, the space between the condenser plates is only partially filled by the material.

Materials which are poor insulators require a special setup; for example, using the method of P. DRUDE: A LECHER line (Sect. 12.4) is

Table 13.1 The dielectric constant ε of various materials

Liquid air	1.5
Petroleum	2
Amber	2.8
Bakelite board	4
Polyvinyl chloride (PVC) (50 Hz)	3.4
Porcelain	4–6
Glasses	6–8

immersed in the liquid to be measured and the reduction of the wavelength of radiation along it relative to its wavelength in air (vacuum) is determined at that frequency (see also Sect. 13.11 and Fig. 12.12).

Some results of measurements for various dielectric materials can be found in Table 13.1.

13.3 Two Quantities Which Can Be Derived from The Dielectric Constant ε

The dielectric constant, which is in general readily measured, yields two other quantities which are important for physics and chemistry. They are:

1. The *dielectric polarization*:[C13.2]

$$P = D_\text{m} - D_0\,. \tag{13.3}$$

The dielectric polarization is thus the additional contribution to the field D_m which results from production or orientation of *electric dipoles* in matter (Fig. 2.56). Its unit is e.g. A s/m^2. Another definition is equivalent: The dielectric polarization of a material is:

$$P = \frac{\text{Electric dipole moment}\, p}{\text{Volume}\, V}\,. \tag{13.4}$$

Derivation: We imagine a block of material (base area A, height l) to be homogeneously polarized. Then on its end surfaces, there is a *bound electric charge*[C13.3] $Q = PA$. Furthermore, from Sect. 3.9, its electric dipole moment $p = Q\,l = PA\,l = PV$, and therefore, $P = p/V$.

2. The *dielectric susceptibility*

$$\chi_\text{e} = \varepsilon - 1 = \frac{D_\text{m} - D_0}{D_0} = \frac{P}{\varepsilon_0 E_0}\,. \tag{13.5}$$

The susceptibility can also be referred to the density ϱ of the material, χ_e/ϱ. This quantity is called the *relative susceptibility* of the material.

C13.2. It is not necessary to use vector notation at this point, while we are considering a parallel-plate condenser; it will be introduced later in Sect. 13.5.

C13.3. In contrast to the "free" charges on the plates of the condenser.

13.4 Distinguishing Dielectric, Paraelectric and Ferroelectric Materials

After explaining the measurement methods, we now wish to give a brief overview of the behavior of insulating materials ("dielectrics") in an electric field. All these materials can be grouped into three large classes, which can be considered to be ideal limiting cases:

1. *Dielectric materials with non-polar molecules*

Their dielectric constants ε and susceptibilities $\chi_e = \varepsilon - 1$ are *material constants*, i.e. they are independent of the electric field which produced the polarization. At constant density, they are also independent of the temperature.

In an atomic picture, this means that the molecules of these dielectric materials themselves have no electric dipole moments. Their dipole moments are *induced* by the applied electric field due to influence (Fig. 2.56). The non-polar molecules are *electrically deformed* by the field and thus become *polarized*. Table 13.2 sets out some numerical values for dielectric materials.

2. *Paraelectric materials with polar molecules*

For this group of materials, also, the dielectric constants ε and the susceptibilities χ_e are quantities which are independent of the strength of the electric field; they are material constants. Numerical examples of ε are set out in the second column of Table 13.3. ε and χ_e are however *temperature dependent*; they increase with decreasing temperature.

Interpretation: The molecules of paraelectric materials are not only electrically deformable, like the non-polar molecules of dielectric materials, but also have *permanent electric dipole moments* p_p, independently of the applied field (*polar* molecules; examples are given in Table 13.3). The applied electric field tends to align these initially randomly-oriented microscopic dipoles along its axis; however, the

Table 13.2 The dielectric constants ε of some dielectric materials (ε always $>$ 1 !) (1 atm $= 1.013 \cdot 10^5$ Pa)

Substance	ε
Helium, 1 atm, 20 °C	1.00006
Air, 1 atm, 18 °C	1.00055
Air, 100 atm, 0 °C	1.05404
Carbon dioxide, 1 atm, 0 °C	1.00095
Bromine vapor, 0.16 atm, 20 °C	1.00035
O_2 liquid, −183 °C	1.464

Table 13.3 Dielectric constants ε and molecular electric dipole moments p_p of some paraelectric materials (cgs unit: 1 Debye $\widehat{\approx}$ $3.4 \cdot 10^{-30}$ A s m) (**Exercises 13.6, 13.7**)

Substance		ε	p_p in A s m
Ammonia (NH_3) (gas)	1 atm, 0 °C	1.0072	$6.13 \cdot 10^{-30}$
KCl (in a molecular beam)	—	—	$34 \cdot 10^{-30}$
Ice	−20 °C	16	—
Methyl alcohol	18 °C	31.2	$5.60 \cdot 10^{-30}$
Glycerine	18 °C	56.2	—
Water	18 °C	81.1	$6.1 \cdot 10^{-30}$
HCl	1 atm, 0 °C	—	$3.4 \cdot 10^{-30}$

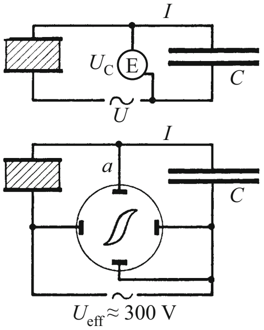


Figure 13.3 The influence of the voltage U on the dielectric constant ε of a ferroelectric crystal. *Above*: Schematic of the principle; *Below*: Observation using an oscilloscope (at the left is a block of SEIGNETTE salt crystal with a thickness of $d = 1$ cm, and two end surfaces of ca. $3 \times 3\,\mathrm{cm}^2$. The right-hand condenser has a capacitance of around $2 \cdot 10^{-6}$ F, the left-hand condenser about $5 \cdot 10^{-8}$ F; therefore, nearly all of the voltage drop U occurs across the crystal).

molecular thermal motions oppose the alignment and tend to maintain the random orientations of the molecules (see Sect. 13.10).

3. *Ferroelectric materials*

As representatives of this group, we mention potassium sodium tartrate ($C_4H_4O_6KNa + 4\,H_2O$), discovered by the pharmacist P. SEIGNETTE (1660–1719); and barium titanate ($BaTiO_3$), as well as potassium dihydrogen phosphate (KH_2PO_4, abbreviated KDP).

Ferroelectric materials are characterized by the extraordinarily large magnitudes attained by their dielectric constants. Values of several 10^4 have been observed. These values are however not even approximately constant. They depend not only on the applied field strength ($\sim U$), but also on the previous history of the sample.

For a demonstration, the setup sketched in Fig. 13.3 (top) is in principle suitable. It consists of two condensers in series, connected to an alternating current source. The capacitance of the right-hand condenser is very large compared to that of the left-hand condenser. As a result, for a given voltage U, the current I is practically determined by the capacitance of the left-hand condenser alone; it is proportional to the voltage U of the current source and to the dielectric constant ε of the material in the left-hand condenser, i.e. $I \sim \varepsilon U$. This current produces the voltage U_C between the plates of the right-hand condenser; it is likewise proportional to I, so that $\varepsilon U \sim U_\mathrm{C}$.

We now want to show experimentally how εU depends on U. This can be most simply accomplished by using an oscilloscope (Fig. 13.3, bottom). Figure 13.4 shows an example: It is a complicated curve,

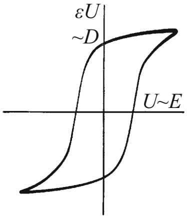


Figure 13.4 A hysteresis loop, registered with the setup shown in Fig. 13.3 (the coordinate axes were drawn in after the measurement). The ordinate is proportional to εU or D; the abscissa is proportional to U or E. D is the displacement density and $E = U/d$ is the electric field strength in the crystal.

called a *hysteresis loop* (Vol. 1, Fig. 8.14). Corresponding pairs of values on the ordinate and the abscissa show that the quantity ε is by no means constant.

Above a certain temperature (the CURIE temperature, for the SEIGNETTE salt about 25°C), the hysteresis loop degenerates into a straight line which has only a small slope relative to the abscissa: Thus, ε is small and constant above the CURIE temperature; there, one finds only the normal behavior as seen in the materials of the 1st and 2nd groups.[C13.4]

C13.4. See for example F. Jona and G. Shirane, "Ferroelectric Crystals", Pergamon Press, 1962. At this point, we should also again mention the *electrets* which were already described in Sect. 3.10. These are materials with a "frozen in" polarization P. They correspond to permanent magnets made of ferromagnetic substances, in the case of *magnetic* dipoles. For technical applications (e.g. microphones and loudspeakers), electrets are usually applied in the form of thin films (~10 µm thick).

13.5 Definitions of the Electric Field Quantities *E* and *D* Within Matter

The definitions given in Sects. 13.1 and 13.3 for the material constants of dielectrics are straightforward and correct, but they refer only to the average values as determined by the condenser setups used there. We now want to give general definitions of the field quantities $\boldsymbol{E}$ and $\boldsymbol{D}$ in the *interior* of a dielectric, based on experimental results:

Let us consider a parallel-plate condenser (e.g. as in Fig. 13.1), which is initially completely filled with a homogeneous dielectric material. A section of it is sketched in Fig. 13.5; it can be seen to contain two small cavities. One of them is a flat *transverse slit*, perpendicular to the direction of the applied field[1]; the other is an *axial channel* which is parallel to the field direction. Both cavities serve to accommodate a measuring instrument (real or virtual), indicated by a dot in the figure. In both cavities, the fields $\boldsymbol{E}$ and $\boldsymbol{D}$ are measured. We find that they are different in the two cavities. In detail, the results are as follows:

1. The field $\boldsymbol{D}$ measured in the *transverse slit* has the same magnitude as that measured in Fig. 13.1 as condenser charge/condenser plate area, i.e. $D = D_{\mathrm{m}}$. This is easily understood. Imagine that the cavity were located directly adjacent to one of the condenser plates. In the limit of vanishing thickness, we define this field as the *displacement density within matter*, and denote it by the vector $\boldsymbol{D}$.[C13.5]

C13.5. The field $\boldsymbol{D}$ is thus determined here by the charges on the condenser plates, the so-called free charges. In contrast, the field $\boldsymbol{E}$ is determined by *all* the charges (free and bound). The MAXWELL equation (Eq. (6.21)) which holds in vacuum can be correspondingly rewritten in the presence of matter as:

$$\operatorname{div} \boldsymbol{E} = \frac{1}{\varepsilon_0}(\varrho_{\text{free}} + \varrho_{\text{bound}})\,.$$

2. The electric field $\boldsymbol{E}$ found in the *axial channel* has the same magnitude as that measured in the empty condenser, i.e. $E = E_0$. This can be shown for a channel of suitable dimensions (Fig. 13.6): The part of the condenser plate which lies above the channel is separated from the rest of the plate by a slit. We measure the displacement density Q/A within this channel and find that it is the same as in the empty condenser, independently of the width of the channel. This result can be extended in a thought experiment to all axial channels, even if they

[1] The direction of the field in a parallel-plate condenser is always well defined. In other cases, we can imagine a sufficiently small spherical cavity within the material. The field direction found within the cavity is identified as the field present *within the matter*.

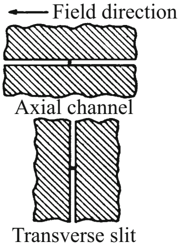


Figure 13.5 The geometry of the cavities: An axial channel and a transverse slit

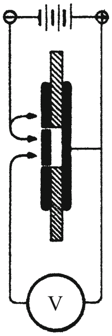


Figure 13.6 The equality of the electric field $\boldsymbol{E}$ in matter and in an axial channel

are too narrow to allow direct measurements. We then obtain E in the cavity by dividing this displacement density by the field constant ε_0. As a result of the equality of E and E_0, the relation $\int E\mathrm{d}s = U$ holds also for E (Eq. (2.3)). For this reason, in the limit of vanishing thickness of the axial channel, we define the quantity E as the *electric field within matter* and denote it by the vector $\boldsymbol{E}$.

By introducing these vector fields $\boldsymbol{E}$ and $\boldsymbol{D}$ within matter, we arrive with the aid of Eq. (13.3) at the definition of the vector field $\boldsymbol{P}$, the electric polarization:C13.6

$$\boldsymbol{P} = \boldsymbol{D} - \varepsilon_0 \boldsymbol{E}\,. \qquad (13.6)$$

In materials with a constant value of ε (and thus no hysteresis), it follows from Eqns. (13.1) and (13.5) that the relations

$$\boldsymbol{D} = \varepsilon\varepsilon_0 \boldsymbol{E} \qquad (13.7)$$

and

$$\boldsymbol{P} = \varepsilon_0(\varepsilon - 1)\boldsymbol{E} = \chi_\mathrm{e}\varepsilon_0 \boldsymbol{E} \qquad (13.8)$$

hold. These three equations describe the relations between the vector fields at each point in space.

C13.6. Equation (13.6) demonstrates in an especially clear way the difference between the fields $\boldsymbol{E}$ and $\boldsymbol{D}$ in matter which we already pointed out in Comment C2.13. Think for example of the fields within a uniformly polarized disk of matter (an electret) (see also Sect. 13.6).

Note that the direction of the vector $\boldsymbol{P}$ is defined in such a way that it points from the negative (bound) charges to the positive (bound) charges (just like the dipole moment $\boldsymbol{p}$ of an electric dipole, Sect. 3.9).

13.6 Depolarization

The vector fields introduced for the interior of dielectric matter will now be investigated in the case that a piece of matter is brought into

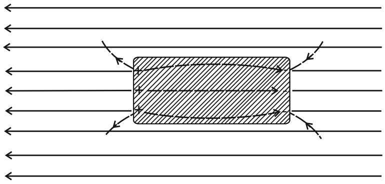


Figure 13.7 Depolarization. The field produced by the influence charges is superposed on the fixed, homogeneous external field. Thus, the total field within the matter is reduced. The total field in the space outside the matter is also modified by this superposition.

a constant, fixed external field $\boldsymbol{E}_0$[2] (Fig. 13.7). This field is presumed to be homogeneous and produced by charges which are located some distance away, so that they are not influenced by the polarization of the piece of matter, i.e. their spatial arrangement remains unchanged (**Exercise 13.5**).

We begin with a simple case, an extended disk with its plane oriented perpendicular to the field $\boldsymbol{E}_0$. According to Sect. 13.5, it follows that the displacement density $\boldsymbol{D}_0 = \varepsilon_0 \boldsymbol{E}_0$ directly in front of the disk is equal to the field $\boldsymbol{D}$ in its interior, $\boldsymbol{D} = \varepsilon\varepsilon_0 \boldsymbol{E}$. Then we find

$$\boldsymbol{E} = \frac{1}{\varepsilon}\boldsymbol{E}_0 \,. \tag{13.9}$$

Making use of Eq. (13.8), we can cast the above equation in the form

$$\boldsymbol{E} = \boldsymbol{E}_0 - \frac{1}{\varepsilon_0}\boldsymbol{P} \,. \tag{13.10}$$

Thus, due to the polarization, the electric field $\boldsymbol{E}$ within the piece of matter is smaller than $\boldsymbol{E}_0$ ($\boldsymbol{E}_0$ and $\boldsymbol{P}$ are parallel). This is called *depolarization*.

In a second example, instead of the disk, a long dielectric rod is brought into the external field, parallel to $\boldsymbol{E}_0$. It then follows, again as in Sect. 13.5, that

$$\boldsymbol{E} = \boldsymbol{E}_0 \,. \tag{13.11}$$

This is easy to understand, since the polarization charges at the ends of the rod are far apart and the surface area of the ends is small. The depolarization effect is thus negligible.

With a sample in the shape indicated in Fig. 13.7, we could suppose that the depolarization would have a value intermediate between these two examples. Indeed, within an ellipsoid of rotation whose

[2] Not to be confused with the field which we have thus far denoted as E_0, which was the field of an empty condenser, produced by the charges on its plates, and which would be changed on inserting matter into the condenser due to the additional charges which would then flow onto the plates.

Table 13.4 Depolarization or demagnetization factors N for rotational ellipsoids

$\frac{\text{Length}}{\text{Diameter}}$	0 (Disk)	1 (Sphere)	0.1	0.2	10	20	50	∞ (infinite wire)
	1	$\frac{1}{3}$	0.863	0.77	0.0203	0.0068	0.0014	0

axis of rotational symmetry is parallel to $\boldsymbol{E}_0$, the depolarization effect leads to a homogeneous field $\boldsymbol{E}$, which is likewise parallel to $\boldsymbol{E}_0$. Then we find[C13.7]

$$\boldsymbol{E} = \boldsymbol{E}_0 - \frac{N}{\varepsilon_0}\boldsymbol{P}\,, \tag{13.12}$$

where $\boldsymbol{P}$ is also homogeneous and parallel to $\boldsymbol{E}_0$. N, called the *depolarizing factor*, is a number which is determined by the ratio of the length of the axis of rotation to the diameter of the ellipsoid; see Table 13.4. With Eq. (13.8), we can rewrite Eq. (13.12):

$$\boldsymbol{E} = \frac{1}{1 + N(\varepsilon - 1)}\boldsymbol{E}_0\,. \tag{13.13}$$

Equations (13.9) and (13.11) represent special cases of this equation, with $N = 1$ (disk) and $N = 0$ (rod).[C13.8]

In the special case of a spherical sample, $N = \frac{1}{3}$ and thus in its interior, the field strength is only

$$\boldsymbol{E} = \frac{3}{(\varepsilon + 2)}\boldsymbol{E}_0\,. \tag{13.14}$$

This likewise homogeneous field can be considered to be the sum of the external field $\boldsymbol{E}_0$ and the field produced by a uniformly polarized sphere, $\boldsymbol{E}_\text{p}$,

$$\boldsymbol{E} = \boldsymbol{E}_0 + \boldsymbol{E}_\text{p}\,. \tag{13.15}$$

Comparison with Eq. (13.12) yields

$$\boldsymbol{E}_\text{p} = -\frac{1}{3\varepsilon_0}\boldsymbol{P}\,. \tag{13.16}$$

The field $\boldsymbol{E}_\text{p}$ is thus determined only by the polarization. Furthermore, it is independent of the manner in which the polarization was produced (if this were not the case, we could construct a sphere with $\boldsymbol{P} = 0$ and $\boldsymbol{E}_\text{p} \neq 0$!). Then, Eq. (13.16) holds e.g. also for a spherical electret.

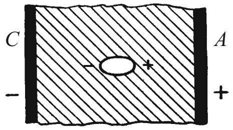
13.7 The Electric Field Within a Cavity

If we were to exchange the roles of the dielectric material and the empty space in Fig. 13.7, as suggested in Fig. 13.8, then we could expect, due to the polarization charges on the inner surfaces of the

C13.7. For a derivation, see for example A. Sommerfield, "*Electrodynamics*" (*Lectures on Theoretical Physics*, Vol. III), Academic Press, 1952, Sect. 13. See also Richard M. Bozorth, "*Ferromagnetism*", D. Van Nostrand Co., Ltd., 1953, Chap. 9.

C13.8. Depolarization also occurs in the experiment of Fig. 13.1; there, $N = 1$, although the field there was held constant using the voltage U. This is however the total field. The "polarizing" field $\boldsymbol{E}_0$, which is produced by the condenser charges (the "free" charges), indeed increases when the dielectric is inserted owing to the charges which then flow onto the condenser plates. This increase in the field is just compensated by the polarization.

Figure 13.8 A cavity in a polarized dielectric



empty space, that the field $\boldsymbol{E}_\text{i}$ would be greater than the field outside. In the case of a spherical body with a dielectric constant ε_i *within* a dielectric material of dielectric constant ε_a, this (homogeneous) field is:C13.9

C13.9. For a derivation, see F. Hund, "*Theoretische Physik*", B.G. Teubner, 3rd ed., 1957, Vol. 2, Sect. 18.

$$\boldsymbol{E}_\text{i} = \frac{3\varepsilon_\text{a}}{\varepsilon_\text{i} + 2\varepsilon_\text{a}} \boldsymbol{E}_\text{a}\,, \tag{13.17}$$

where $\boldsymbol{E}_\text{a}$ is the field of the material in the outer region at some distance from the cavity, i.e. the homogeneous field that would be present in the material without the cavity. For $\varepsilon_\text{a} = 1$, we find Eq. (13.14). For $\varepsilon_\text{i} = 1$ (for example the interior of an empty spherical bubble), it follows that

$$\boldsymbol{E}_\text{i} = \frac{3\varepsilon_\text{a}}{2\varepsilon_\text{a} + 1} \boldsymbol{E}_\text{a}\,. \tag{13.18}$$

13.8 Paraelectric and Dielectric Materials in an Inhomogeneous Electric Field

All paraelectric and dielectric materials are pulled toward regions of greater field strength within an inhomogeneous electric field. This recalls the oldest electrical observation, the attraction of scraps of cloth or paper to charged bodies, e.g. a rubbed amber rod (Fig. 3.19). Depolarization makes a quantitative treatment of this phenomenon rather complicated. It succeeds only for bodies of simple, symmetric shape, for example for the attraction between a *small* insulating *sphere* (volume V) and a large charged sphere (radius r). When the distance between their centers is R, we find for the magnitude of the force

$$F = \frac{6r^2 V \varepsilon_0 (\varepsilon - 1)}{\varepsilon + 2} \cdot \frac{U^2}{R^5}\,. \tag{13.19}$$

The force thus decreases as the fifth power of the distance! Figure 13.9 shows an example.

Derivation of Eq. (13.19): Eqns. (3.28) and (13.4) yield

$$F = p\frac{\partial E_\text{R}}{\partial R} = PV\frac{\partial E_\text{R}}{\partial R}\,. \tag{13.20}$$

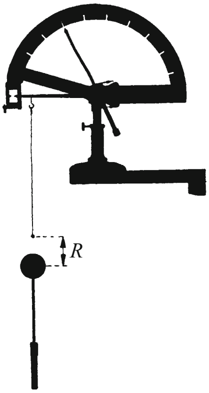


Figure 13.9 The attractive force on a small insulating sphere in the inhomogeneous electric field of a large sphere, measured using a spiral-spring balance (example: an amber sphere of diameter $= 6\,\mathrm{mm}$, $V = 1.13 \cdot 10^{-7}\,\mathrm{m}^3$, $\varepsilon = 2.8$, radius of the charged sphere $r = 2 \cdot 10^{-2}\,\mathrm{m}$, $U = 10^5\,\mathrm{V}$, $R = 5 \cdot 10^{-2}\,\mathrm{m}$, depolarizing factor $N = 1/3$ (Table 13.4), $F = 2.9 \cdot 10^{-5}\,\mathrm{N}$). The reader should look at **Exercise 2.16** in this connection.

At the point of observation, according to Eqns. (2.15) and (2.16), we have

$$E_\mathrm{R} = \frac{Ur}{R^2} \tag{13.21}$$

and

$$\frac{\partial E_\mathrm{R}}{\partial R} = -\frac{2Ur}{R^3}\,. \tag{13.22}$$

The electric polarization of the small sphere is

$$P = \varepsilon_0(\varepsilon - 1)E\,, \tag{13.8}$$

and the field strength in the interior of the sphere is

$$E = \frac{3}{\varepsilon + 2}E_\mathrm{R}\,. \tag{13.14}$$

Combining Eqns. (13.14), (13.20) and (13.22) yields Eq. (13.19).

13.9 The Molecular Electric Polarizability. The Clausius-Mossotti Equation

The different behavior of dielectric and paraelectric materials has already been explained qualitatively in Sect. 13.4. Its *quantitative* explanation is rather important for an understanding of *molecular structure* and thus for *chemistry*. To arrive at it, we need the concept of *molecular electric polarizability*. We derive it using what we have learned about depolarization in the previous sections.

In the interior of a body of volume V, let the electric field be $\boldsymbol{E}$, and assume that it produces an electric polarization in the body:

$$\boldsymbol{P} = \varepsilon_0(\varepsilon - 1)\boldsymbol{E}\,. \tag{13.8}$$

With this polarization, the body acquires an electric dipole moment $\boldsymbol{p}$ parallel to the direction of the field. Then Eq. (13.4) applies. In vector form, it states that

$$\boldsymbol{P} = \frac{1}{V}\boldsymbol{p}\,. \tag{13.23}$$

In an atomic picture, the total electric dipole moment $\boldsymbol{p}$ is interpreted as the vector sum of the average molecular contributions $\boldsymbol{p}'$ which are due to the N individual molecules, that is:

$$\boldsymbol{P} = \frac{N}{V}\boldsymbol{p}' = N_\mathrm{V}\,\boldsymbol{p}' \tag{13.24}$$

($N_\mathrm{V} = N/V =$ number (particle) density of the molecules).

We combine Eqns. (13.8) and (13.24), obtaining:

$$\boldsymbol{p}' = \frac{1}{N_\mathrm{V}}\boldsymbol{P} = \frac{\varepsilon_0(\varepsilon - 1)}{N_\mathrm{V}}\boldsymbol{E}\,. \tag{13.25}$$

Experimentally, we find that ε is constant, and thus the average contribution $\boldsymbol{p}'$ is proportional to the magnitude of the field acting on the molecules, which we call $\boldsymbol{E}_\mathrm{w}$. For this reason, we set

$$\boldsymbol{p}' = \alpha\boldsymbol{E}_\mathrm{w} \tag{13.26}$$

and call α the *molecular electric polarizability*. Note that $\boldsymbol{p}'$, $\boldsymbol{E}_\mathrm{w}$, and α are all *microscopic* quantities, acting at the level of atoms or molecules.

As the effective field $\boldsymbol{E}_\mathrm{w}$, for gases, vapors and dilute solutions, we can simply take the field $\boldsymbol{E}$ which occurs in Eq. (13.25). We then obtain

$$\alpha = \frac{\varepsilon_0(\varepsilon - 1)}{N_\mathrm{V}} \tag{13.27}$$

(e.g. α in $\mathrm{A\,s\,m^2/V}$, $\varepsilon_0 = 8.86 \cdot 10^{-12}\,\mathrm{A\,s/(V\,m)}$).

In liquids and in solid bodies, setting $\boldsymbol{E}_\mathrm{w}$ and $\boldsymbol{E}$ equal is no longer permissable. In these "condensed phases", the molecules are too densely packed, and therefore, their interactions with one another must be taken into account in polarized liquids and solids. This is done in the equation for the molecular electric polarizability that was suggested by CLAUSIUS and MOSSOTTI:

$$\alpha = \frac{3\varepsilon_0}{N_\mathrm{V}} \cdot \frac{\varepsilon - 1}{\varepsilon + 2} \tag{13.28}$$

(for $\varepsilon \approx 1$, this equation reverts to Eq. (13.27)).

Derivation of Eq. (13.28): We start from Eqns. (13.24) and (13.26). They give

$$\boldsymbol{P} = \alpha N_\mathrm{V}\boldsymbol{E}_\mathrm{w}\,. \tag{13.29}$$

To calculate the effective field $\boldsymbol{E}_\mathrm{w}$, we consider a single molecule a. The remaining molecules are divided into two groups of unequal size. In the first, smaller group, we include all those molecules in the immediate neighborhood of molecule a. The boundary of this nearby region is arbitrarily chosen to be a *spherical surface* with a at its center. The second, larger group includes all the molecules outside this sphere. In amorphous substances and regular crystals, the neighboring molecules within the virtual boundary surface around molecule a are arranged with spherical symmetry. Therefore, their influence cancels mutually, and only the effect of the second group remains. The molecule a floats, figuratively speaking, within a spherical "cavity" inside a *homogeneously* polarized body. To determine $\boldsymbol{E}_\mathrm{w}$, we cannot use Eq. (13.18), since it was derived under the assumption that $\varepsilon = 1$ within the cavity. Then, however, the electric field in the neighborhood of the cavity would not be constant, in contradiction to the case we are dealing with here. We therefore assume that at the position of molecule a, the field $\boldsymbol{E}$ can be expressed as the sum of the field of a uniformly-polarized sphere (Sect. 13.6) and the field we are seeking, $\boldsymbol{E}_\mathrm{w}$:

$$\boldsymbol{E} = \boldsymbol{E}_\mathrm{w} + \boldsymbol{E}_\mathrm{sphere}\,, \tag{13.30}$$

where $\boldsymbol{E}$ is the field in a uniformly-polarized body (without a cavity) and $\boldsymbol{E}_\mathrm{sphere}$ is the field in a sphere with the same polarization, $\boldsymbol{P}$ (see Eq. (13.16)); then

$$\boldsymbol{E}_\mathrm{w} = \boldsymbol{E} + \frac{1}{3\varepsilon_0}\boldsymbol{P}\,. \tag{13.31}$$

Then, using Eq. (13.8), we find:C13.10

$$\boldsymbol{E}_\mathrm{w} = \frac{\varepsilon + 2}{3}\boldsymbol{E}\,, \tag{13.32}$$

and from this, using Eqns. (13.25) and (13.26), we finally obtain Eq. (13.28).

C13.10. Thus, the internal field $\boldsymbol{E}$ that we defined in Sect. 13.5 does *not* act on the single molecule within the dielectric material, but instead the field $\boldsymbol{E}_\mathrm{w}$ acts there. It – depending on the value of the dielectric constant ε (Table 13.1) – may be considerably larger!

The two equations (13.27) and (13.28) permit a relatively simple determination of the molecular electric polarizability α: We need only *measure* the dielectric constant ε and insert it, together with the relevant number (particle) density N_V. Table 13.5 contains some numerical values for the polarizability α of *non-polar* molecules in liquids (**Exercise 13.6**).

Table 13.5 Electric polarizabilities of some non-polar molecules in liquids (at ≈ 20 °C) (N_A is the AVOGADRO constant, $6.022 \cdot 10^{23}\ \mathrm{mol}^{-1}$)

Substance	Molar mass $M_\mathrm{n} = \dfrac{M}{n}$	Density ϱ	Number density of the molecules $N_\mathrm{V} = \dfrac{\varrho N_\mathrm{A}}{M_\mathrm{n}}$	Dielectric constant ε	Electric polarizability α
	in $\dfrac{\mathrm{kg}}{\mathrm{kmol}}$	in $\dfrac{\mathrm{kg}}{\mathrm{m}^3}$	in m^{-3}		in $\dfrac{\mathrm{A\,s\,m}^2}{\mathrm{V}}$
Carbon disulfide, CS_2	76	1250	$9.9 \cdot 10^{27}$	2.61	$0.94 \cdot 10^{-39}$
Biphenyl, C_6H_5–C_6H_5	154	1120	$4.37 \cdot 10^{27}$	2.57	$2.1 \cdot 10^{-39}$
Hexane, C_6H_{14}	86	662	$4.63 \cdot 10^{27}$	1.88	$1.3 \cdot 10^{-39}$

13.10 The Permanent Electric Dipole Moments of Polar Molecules

In paraelectric materials, the measurements show a decrease in the molecular electric polarizability with increasing temperature. Figure 13.10 shows a typical example for a gas, HCl. We can recognize two different contributions, one which is independent of the temperature (the thin horizontal line below), and the other which depends on the temperature (above the constant part).

Interpretation: The temperature-*in*dependent part is caused by an electrical deformation of the molecules, as is well known for dielectric materials (see Sect. 2.17), and was explained in Fig. 2.56. The temperature-dependent part adds to it, and is due to the fact that the molecules of paraelectric materials have permanent electric dipole moments p_p even in the absence of an external electric field.

Without a field, the orientations of p_p are randomly distributed owing to the thermal motions of the molecules. The sum of the electric dipole moments p_p, averaged spatially and over time, is equal to zero. An electric field however provides a preferred axis of orientation for the dipole moments p_p. Each molecule then has a net component along the field axis, averaged over time, and this results in an average moment p' as the contribution of a single molecule. The magnitude of p' is only a fraction (usually a very small fraction) x of the permanent dipole moment p_p, that is

$$p' = x p_\mathrm{p} \,. \tag{13.33}$$

The fraction x is given by

$$x \approx \frac{1}{3}\frac{p_\mathrm{p} E_\mathrm{w}}{kT} \tag{13.34}$$

($k =$ BOLTZMANN constant $= 1.38 \cdot 10^{-23}$ W s/K).

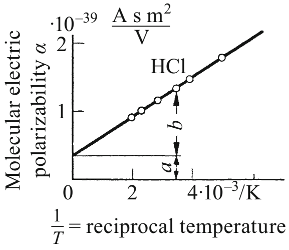


Figure 13.10 The polarizability of a dipolar molecule as defined by Eq. (13.26), as a function of temperature. The constant part a is due to "influence" or "molecular deformation", while the temperature-dependent part b is due to alignment of the thermally disordered polar molecules with their permanent dipole moments. Only this second part should be inserted into Eq. (13.35) (at a pressure of 1 atm ($= 1.013 \cdot 10^5$ Pa); measurement frequency $\nu = 1$ MHz) (C.T. Hahn, *Physical Review* **24**, p. 400 (1924)).

The fraction x is thus essentially equal to the ratio of two energies: (i) the work $\boldsymbol{p}_{\mathrm{p}} \cdot \boldsymbol{E}_{\mathrm{w}}$ would be required in order to align the dipole *perpendicular* to the field axis; (ii) kT is the thermal energy which is on the average transferred by molecular collisions (and tends to disorient the dipoles). A precise calculation would have to take not only the perpendicular alignment into account, but also all other possible orientations by computing the appropriate average (LANGEVIN-DEBYE formula). This leads in the first approximation to the numerical factor of 1/3.

We combine Eqns. (13.33) and (13.34) with Eq. (13.26), denoting the temperature-dependent part of α in Eq. (13.26) as α_{T}, and obtain for the *permanent electric dipole moment of the dipolar molecules*:

$$p_{\mathrm{p}} \approx \sqrt{\alpha_{\mathrm{T}} 3kT}\,. \qquad (13.35)$$

For the example of the HCl molecule, the measurements in Fig. 13.10 give the molecular electric polarizability at 273 K:

$$\alpha_{\mathrm{T}} = 1.05 \cdot 10^{-39}\,\frac{\mathrm{A\,s\,m^2}}{\mathrm{V}}\,.$$

Inserting this value into Eq. (13.35), we obtain for the permanent electric dipole moment of a single HCl molecule, $p_{\mathrm{p}} \approx 3.4 \cdot 10^{-30}$ A s m (cf. Table 13.3) (**Exercise 13.7**).

One could therefore imagine the molecule from an electrical point of view as consisting of two electrical elementary charges, each of $1.602 \cdot 10^{-19}$ A s, at a spacing of around $0.2 \cdot 10^{-10}$ m. (For comparison: The order of magnitude of the molecular diameter is 10^{-10} m.)

Making use of this value of p_{p}, we can calculate the fraction x in Eq. (13.34). The field strength may be large, namely $E = 10^6$ V/m, and the temperature $T = 300$ K. Then we find $x = 3 \cdot 10^{-4}$, that is $x \ll 1$. Thus, the average contribution p' of the permanent dipole moment p_{p} to the electric polarization P is still proportional to the field strength, and the susceptibility, $P/\varepsilon_0 E = \chi_{\mathrm{e}} = (\varepsilon - 1)$, is constant (linear region). Only at very low temperatures can x approach a value of 1 with increasing field strength, causing the electric polarization to saturate at its maximum possible value.

13.11 The Frequency Dependence of the Dielectric Constant ε

In order to determine the capacitance $C = Q/U$, we measure the charge Q which is stored in a condenser at the voltage U. In doing this, we make a tacit assumption: The magnitude of the charge Q stored at a given voltage U is supposed to be time-independent. This can however not be generally true. The processes which occur within

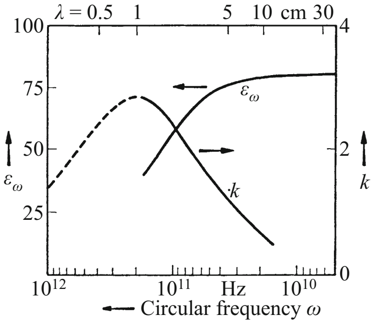


Figure 13.11 The frequency dependence of the dielectric constant and the absorption coefficient k of water. The corresponding vacuum wavelengths are shown on the abscissa above (temperature 18 °C)

the dielectric after the field is applied are not instantaneous; rather, they require a certain time on the average. After one "relaxation time" τ_r, a fraction $1/e \approx 37\,\%$ is still missing from the equilibrium values. Normally, the time during which the field remains constant is much longer than the relaxation time τ_r. However, if it becomes comparable to τ_r, then the measured (apparent) dielectric constant ε is reduced, and at each circular frequency ω, we would obtain a different value of the dielectric "constant", ε_ω.C13.11

C13.11. Another cause of a frequency dependence of the measured dielectric constant is discussed in Chap. 27, under the topic of optics; see in particular Sect. 27.7.

Figure 13.11 shows as an example the dielectric constant ε_ω of water at 18 °C in the range of circular frequencies from $\omega = 5 \cdot 10^9$ Hz to $2 \cdot 10^{11}$ Hz, that is for fields with oscillation periods between 10^{-8} s and 10^{-11} s. For pure water, the relaxation time is $\tau_\text{r} = 3 \cdot 10^{-11}$ s, determined by the rate of alignment of the dipolar molecules in the applied field. This is hindered by a resistance similar to friction, which consumes energy (dissipation). It is greatest around the circular frequency $\omega_\text{R} = 1/\tau_\text{r}$. In this frequency range, electrical waves are strongly absorbed by water (this is the principle of microwave ovens; see **Exercise 13.8**). In Fig. 13.11, the absorption coefficient k, often used in optics (Optics, Sect. 25.3), is also plotted.

The relationship between the dielectric constant and the power dissipation (absorption accompanied by heating) on the one hand, and the relaxation time τ_r on the other, is quite general. It is unimportant precisely which physical processes are responsible for the relaxation. We explain this by referring to a simple model.

In Fig. 13.12, at the left, we show the model system: We imagine the dielectric to consist of layers. The shaded layers are supposed to be perfectly insulating, while the dotted layers are poor conductors, with large electrical resistances. For this layered dielectric, Fig. 13.12 (right) shows an equivalent circuit, the well-known series circuit from Sect. 10.6 with an AC current source (relaxation time $\tau_\text{r} = RC$, Sect. 2.16).
The resistor in the equivalent circuit has two effects: First, the amplitude of an AC current of circular frequency ω is reduced by the factor

$$\frac{I_\text{C,R}}{I_\text{C}} = \frac{\text{Current amplitude with } R}{\text{Current amplitude without } R} = \frac{1}{\sqrt{1 + (\omega\tau_\text{r})^2}} = |\varepsilon_\omega|\,. \quad (13.36)$$

(This quantity is called the relative dielectric constant. Equation (13.36) follows from Eqns. (10.29) and (10.30)). Second, the AC current is no

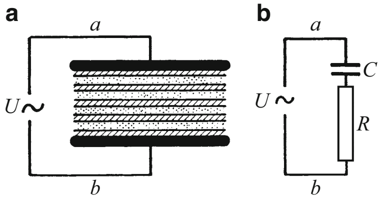


Figure 13.12 *At left*: A layered dielectric. *At right*: Its equivalent circuit

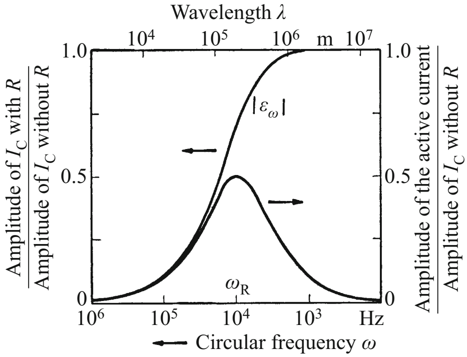


Figure 13.13 The frequency dependence of the relative dielectric constant (Eq. (13.36)) and the loss factor (Eq. (13.39)) in the equivalent circuit of Fig. 13.12 for a relaxation time of $\tau_r = RC = 10^{-4}$ s (λ is the wavelength of electromagnetic radiation at the circular frequency ω)

longer phase-shifted by 90° relative to the voltage, but rather only by a smaller phase angle φ. From Eq. (10.31), we find

$$\tan\varphi = -\frac{1}{\omega\tau_r} \tag{13.37}$$

(the current leads the voltage; cf. Eq. (10.24)).

In addition to the *reactive* current (or "idle" current), there is also an *active* current of amplitude[C13.12]

$$I_{act} = U\omega C\,\frac{\omega\tau_r}{1+\omega^2\tau_r^2}\,. \tag{13.38}$$

The ratio of the active current amplitude to the total current amplitude in a loss-free condenser is called the *loss factor*, and it is found to be

$$\varepsilon_{\omega,act} = \frac{\omega\tau_r}{1+\omega^2\tau_r^2}\,. \tag{13.39}$$

The results of Eqns. (13.36) and (13.39) are shown graphically in Fig. 13.13: At a frequency of $\omega_R = 1/\tau_r$, the loss factor $\varepsilon_{\omega,act}$ exhibits a maximum.

C13.12. In deriving Eq. (13.38), note that:

$I_{act} = I_{C,R}\cos\varphi\,.$

(see Sect. 10.8)

Exercises

13.1 A dielectric slab of dielectric constant ε is slid into a charged parallel-plate condenser with a charge of Q, so that it fills the entire volume between the condenser plates. How does the energy stored in the condenser change, and where does the energy difference go? (Sect. 13.2)

13.2 A sheet of glass of thickness $d < l$ is slid between the condenser plates of a charged parallel-plate condenser (the spacing of the plates is l). This causes its voltage to drop from U to U', similar to what is shown in Video 2.1. Find the dielectric constant ε of the glass, and compute it for the case that $d = 0.5\,l$ and $U' = 0.6\,U$.

13.3 How is the bridge circuit in Fig. 13.2 used to determine the capacitance C_x? (Sect. 13.2)

13.4 A long dielectric rod with a dielectric constant of ε is placed in an electric field E with its long axis parallel to the field direction. How large is the displacement density D in the rod? (Sect. 13.6)

13.5 Find the polarization P of a dielectric sphere with a dielectric constant of ε in a homogeneous electric field E_0. (Sect. 13.6)

13.6 Make use of the CLAUSIUS-MOSSOTTI equation (13.28) to find the dipole moment p_p of liquid water using the data given in Table 13.3, and compare the result with the correct value, which is also given in Table 13.3 (it however was determined for the gas phase). Apparently, the CLAUSIUS-MOSSOTTI equation cannot be applied to water in the liquid phase. Why not? (Sect. 13.9)

13.7 In the reference work *Landolt-Börnstein*, 6th ed. (Springer, Berlin 1959), Vol. II, Part 6, p. 874, we find for the dielectric constant of ammonia (NH_3) at a temperature of 22.5 °C and a pressure of 1 atm ($= 1.013 \cdot 10^5$ Pa) a value of $\varepsilon = 1.00612$. Explain the difference compared to the value in Table 13.3. (Sect. 13.10)

13.8 In a microwave oven (frequency $\nu = 2.5$ GHz), a liter of water is heated by 10 °C in one minute.
a) Find the wavelength λ_W of the microwave radiation in water. b) Determine the average power $\overline{\dot{W}}$ which is absorbed by the water. c) The water is contained in a cube-shaped plastic beaker which is 10 cm on a side, and the radiation impinges on it from all directions. Compare the radiation power per surface area (i.e. the magnitude of the POYNTING vector $\boldsymbol{S}$, see Comment C12.5) with the solar constant for onto the surface of the Earth, $E_e = 1.37\,\mathrm{kW/m^2}$, Sect. 19.3. Neglect reflection and heat losses. (Sect. 13.11)

Matter in a Magnetic Field 14

14.1 Introduction. The Permeability μ

Thus far, we have considered magnetic fields only in empty space (the presence of air molecules had a negligible effect on our results. The influence of the air changes only the 6th place after the decimal by 4 units).

Some of the current-carrying conductors that we considered, in particular coils, were not self-supporting, but rather were wound on coil forms or spools, e.g. thin-walled cardboard or wooden tubes covered with an insulating layer of shellac. The effects of these spools was also negligible within the precision needed for our demonstration experiments.

In contrast, the presence of some other materials in the magnetic field, e.g. of iron, has a rather drastic effect. When a ring coil is wound around an iron core (Fig. 14.1), its magnetic flux Φ is increased by a large factor (and no field lines are found outside the coil in this case; cf. Fig. 14.2). Based on this fact, we define the *permeability* of the filling material to be the ratio[C14.1]

$$\mu = \frac{\text{Magnetic flux } \Phi_\mathrm{m} \text{ of the filled ring coil}}{\text{Magnetic flux } \Phi_0 \text{ of the empty ring coil}} .$$

Dividing the magnetic flux Φ by the cross-sectional area A of the homogeneous magnetic field, one obtains its *flux density* B, i.e. $B = \Phi/A$. Then for the definition of the *permeability*, we find

$$\mu = \frac{B_\mathrm{m}}{B_0} . \tag{14.1}$$

[C14.1] The permeability μ is often also denoted in the literature by μ_r (r stands for "relative to vacuum"); it is given the name "relative permeability". Frequently, the product $\mu_0\mu_\mathrm{r}$ is simply denoted by μ, and called "the permeability". μ_0 is thus the "permeability of vacuum".

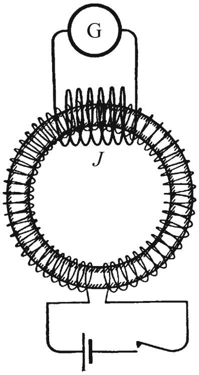


Figure 14.1 Defining magnetic material constants by measurements of the magnetic flux density. For demonstration experiments, the ring coil can also be filled with "Ferrocart", an iron-containing paper-based material with a permeability μ of about 10. (See Sect. 5.4)

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_14

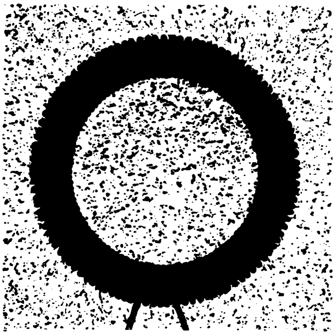


Figure 14.2 The outside space around a ring coil (toroidal coil) with an iron core is field-free (this holds also for a ring coil without an iron core; see Fig. 4.8)

14.2 Two Quantities Derived from the Permeability

Two other, often-used physical quantities are defined with reference to the permeability μ:

1. The *magnetization* M is defined by the equation[C14.2]

$$\mu_0 M = B_\mathrm{m} - B_0 = (\mu - 1) B_0 \, . \tag{14.2}$$

We thus define the magnetization M as an additional contribution to the magnetic flux density in matter, which is in turn produced by the applied field. $\mu_0 M$ is also called the *magnetic polarization*. The unit of M is for example 1 A/m. A different definition is also equivalent: It refers to the magnetic moment within a material:

$$M = \frac{\text{Magnetic moment } m}{\text{Volume } V} \, . \tag{14.3}$$

Derivation: We imagine a block of matter with a base area A, which is homogeneously magnetized along its length l. Then the magnetic flux Φ which the magnetization M of the block produces is given by $\Phi = (B_\mathrm{m} - B_0)A = \mu_0 MA$; and from Eq. (8.20), its magnetic moment is $m = \Phi \, l/\mu_0 = MA \, l = MV$.

2. The *magnetic susceptibility*

$$\chi_\mathrm{m} = \mu - 1 = \frac{\mu_0 M}{B_0} = \frac{M}{H_0} \, . \tag{14.4}$$

Sometimes the susceptibility is also referred to the density ϱ of the corresponding substance, χ_m/ϱ, or to the amount-of-substance density, $\chi_\mathrm{m}/(n/V)$.[C14.3] Some examples can be found in Table 14.1.

C14.2. The magnetization M is, like the field B, a vector. Since we assume here the simple but frequently-occurring case that M and B are collinear, the vector notation is not necessary at this point. However, compare Sect. 14.5.

C14.3. The magnetic moment m and the magnetization M – and therefore the susceptibility χ_m of a sample of volume V, which can be computed from them – are influenced by the *filling factor*, i.e. whether the material is porous, gaseous etc. For this reason, the susceptibility is often referred to the mass density ϱ; i.e. instead of simply quoting χ_m, the quantity χ_m/ϱ is given. Referring the susceptibility to the amount-of-substance density, $\chi_\mathrm{m}/(n/V) = \chi_m \cdot \frac{V}{n}$, where n is the amount of substance (unit: mol) is also convenient.
(In the cgs unit system, this quantity is often (and somewhat incorrectly) called the “molar susceptibility”. A note on expressing it in cgs units: $\chi_\mathrm{m,SI} \mathrel{\hat{=}} 4\pi \ \chi_\mathrm{m,cgs}$).

14.3 Measuring the Permeability μ

There are many setups suitable for measuring the permeability. If the sample is available in the form of a ring (torus), then the scheme illustrated in Fig. 14.1 may be used. In carrying out the measurements, one can if necessary increase the sensitivity by several orders of magnitude by applying a difference method. The induction coils of the empty and the filled ring coil are connected in opposition, and the difference between the two magnetic fluxes is measured directly as a voltage impulse. An example is illustrated in Fig. 14.6.

For many substances, $\mu \approx 1$. In that case, one does not try to measure μ directly, but determines instead the magnetization M. μ can then be calculated using the defining equations as given in Sect. 14.2. Instead of large, ring-shaped samples, one can use small samples of volume V with arbitrary shapes. They are inserted into a magnetic field, and the magnetic moment which is produced by magnetization is measured:

$$m = MV\,.$$

To this end, the magnetic field is made inhomogeneous and then the force acting on the sample along the direction of the field gradient is measured by some method (e.g. Fig. 14.3):

$$F = m\frac{\partial B}{\partial x} = MV\frac{\partial B}{\partial x} \tag{8.18}$$

(e.g. F in N, m in A m^2, $\partial B/\partial x$ in V s/m^3, V in m^3).
The field gradient $\partial B/\partial x$ is measured as described in the text following Eq. (8.18) in Chap. 8.
For the determination of μ, one uses the magnetic flux density B_0 which is present *without* the sample. It is measured using a small induction coil as the average value at the location of the sample. The application of this field in Eq. (14.2) is only approximately correct. It is however permissible if the permeability μ is ≈ 1. We will see the reason for this later from Eq. (14.17) (**Exercise 14.1**).

14.4 Distinguishing Diamagnetic, Paramagnetic and Ferromagnetic Materials

After explaining the measurement procedure, we now want to give an overview of the magnetic properties of materials. All substances can be collected in three large groups:

1. *Diamagnetic materials*.

Their susceptibilities $\chi_\mathrm{m} = (\mu - 1)$ are, like μ, material constants and are *independent* of the strength of the applied magnetizing field. They are also independent of the *temperature* (for a constant density ϱ). The permeability μ is somewhat smaller than 1. Table 14.1 contains some examples.

Table 14.1 Dia- and paramagnetic substances[a] (1 atm $= 1.013 \cdot 10^5$ Pa)

Diamagnetic substances (μ always $<$ 1), $T \approx 293$ K (20 °C)						
	H_2 (1 atm)	Cu	H_2O	NaCl	Bi	
Susceptibility $\chi_\mathrm{m} = (\mu - 1)$	-0.002_2	-9.6	-9.06	-13.9	-165	$\cdot 10^{-6}$
χ_m/ϱ ($\varrho =$ Density)	-25	-1.08	-9.06	-6.5	-16.8	$\cdot 10^{-9}\ \mathrm{m^3/kg}$
$\chi_\mathrm{m}/(n/V)$ ($n =$ Amount of substance)	-0.5	-0.69	-1.63	-3.78	-35.2	$\cdot 10^{-10}\ \mathrm{m^3/mol}$

Paramagnetic substances (μ always $>$ 1), $T \approx 293$ K (20 °C)						
	Al	Pt	O_2 (gas)	O_2 (liquid)	Dy_2S_3	
			(1 atm)	$T = 90.2$ K	$T = 293$ K	
Susceptibility $\chi_\mathrm{m} = (\mu - 1)$	20.7	264	1.88	3470	17 200	$\cdot 10^{-6}$
χ_m/ϱ ($\varrho =$ Density)	7.67	12.3	1300	3020	2839	$\cdot 10^{-9}\ \mathrm{m^3/kg}$
$\chi_\mathrm{m}/(n/V)$ ($n =$ Amount of substance)	2.07	24	416	967	11 960	$\cdot 10^{-10}\ \mathrm{m^3/mol}$

[a] See the *CRC Handbook of Chemistry and Physics*, 84th Edition, 2003/4 (CRC Press, NY).

2. *Paramagnetic materials.*

Their susceptibilities are also practically independent of the strength of the applied magnetizing field. The permeability μ is somewhat larger than 1. Examples are likewise collected in Table 14.1. Sometimes, their susceptibility χ_m decreases with increasing temperature (cf. Fig. 14.5, Parts B and C). In simple limiting cases, CURIE*'s law* holds:

$$\chi_\mathrm{m} = \frac{C_T}{T} \tag{14.5}$$

(C_T is called the CURIE constant; derivation in Sect. 14.8).

3. *Ferromagnetic materials.*

In these materials, the permeability μ is not even approximately a material constant. It depends not only on the strength of the magnetizing field, but also on the magnetic history of the sample and on its structure, e.g. whether it is bulk material or powdered. The magnitude of μ can exceed 1000. With increasing temperature, the permeability decreases. Above a certain critical temperature (the CURIE temperature), ferromagnetism vanishes and the material exhibits only paramagnetic behavior.

So much for the general classification; now some details:

1. *Diamagnetic materials.* They can be identified in a magnetic field even without a quantitative measurement. They are always forced out of the region where the magnetic field strength is strongest; for example, a piece of bismuth will be pushed upwards in the apparatus shown in Fig. 14.3. Explanation: *Atoms or molecules of diamagnetic materials have no permanent magnetic moments. They acquire a moment only in an applied magnetic field, through induction* (WILHELM WEBER, 1852). The induced currents flow in circles, as shown in Fig. 8.16b, opposing the current in the field coil, without losses as

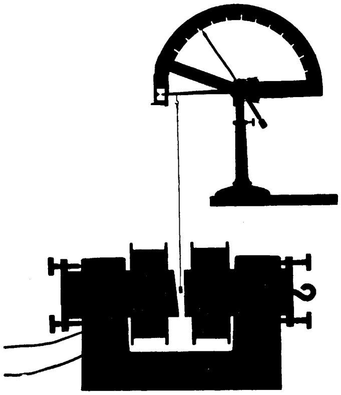


Figure 14.3 A sample in an inhomogeneous magnetic field is hanging from a spiral-spring balance (**Exercise 14.1**)

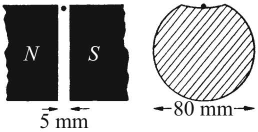


Figure 14.4 Small *diamagnetic* objects ($V \approx 1\,\mathrm{mm}^3$) made of bismuth or well-tempered arclight carbon are levitated on the fringes of a magnetic field with a flux density of $B \approx 2\,\mathrm{V\,s/m^2}$ (an electromagnet as in Fig. 14.3). The concave region milled into the top of the magnet pole piece (radius of curvature 6 cm) stabilizes them in the horizontal direction.

long as the field is present. These induced currents must occur in all atoms, so that *all substances are originally diamagnetic*. Their diamagnetic behavior can however be concealed by other, stronger phenomena; this is the case for para- and ferromagnetic materials.$^{\mathrm{C14.4}}$ Diamagnetic samples can be made to *levitate* in an inhomogeneous magnetic field of a suitable shape (Fig. 14.4), and they will *oscillate* if given an impulse along the direction of a field gradient.$^{\mathrm{C14.5}}$

2. *Paramagnetic materials*. In contrast to diamagnetic materials, they are pulled *into the region of greatest field strength*. Some demonstrations are shown in Figs. 14.5A–C.

Explanation: *The molecules of paramagnetic materials* not only acquire a magnetic moment by induction in a magnetic field, but also they have permanent magnetic moments m_p, independently of the field.$^{\mathrm{C14.6}}$

The dipole axes of these permanent moments are however distributed randomly over all directions as a result of thermal motions. Therefore, the paramagnetic sample exhibits no overall magnetic moment without an applied field. *In an applied magnetic field, however, the atomic moments acquire a preferred direction*. Nevertheless, under usual conditions, the orientation of all the permanent moments parallel to the field is far from complete. The time-averaged contribution m'_p of individual atomic or molecular moments to the total moment m, and thus to the magnetization M, is only a fraction (usually a small

C14.4. Note here the difference between magnetic and electric moments (Chap. 13): Both dielectric and paraelectric materials are always pulled towards regions of stronger fields; that is, χ_e is always *positive*.

C14.5. Superconducting materials show strong diamagnetic behavior due to their expulsion of magnetic fields (MEISSNER-OCHSENFELD effect). They can be used for impressive demonstrations of magnetic levitation. See Comment C10.3.

C14.6. Exception: PAULI paramagnetism, for example in the metals Al and Pt. More details can be found in C. Kittel, *Introduction to Solid State Physics*, 7th ed., Chap. 14 (John Wiley, NY 1996), especially Figs. 1 and 11.

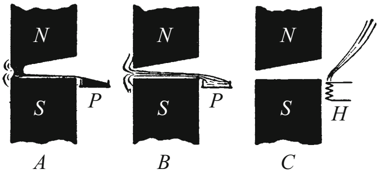


Figure 14.5 A: Liquid oxygen is strongly paramagnetic and is therefore pulled into regions of greater field strength (here, from a cardboard tray P). (The electromagnet used is similar to the one shown in Fig. 14.3). B and C: The temperature dependence of the paramagnetic susceptibility χ_m. Cold air has a larger susceptibility than room air, while warm air has a smaller susceptibility. (Both χ_m/ϱ and the density ϱ are proportional to T^{-1}; χ_m is thus proportional to T^{-2}). As a result, a cloud of cold air is pulled into the region of greater magnetic field strength (from the cooled paper tray P), where it displaces the room air (Part B) **(Video 14.1)**. In contrast, rising warm air (from a flame or a heating coil H) is forced away from the magnet into a region of smaller field strength by the more strongly attracted room air (Part C).

Video 14.1:
"Paramagnetic materials"
http://tiny.cc/kdggoy
The video shows the experiments from Parts A and B of the figure.

fraction) of the permanent molecular moment m_p. Otherwise, the magnetization M in strong applied fields would no longer be proportional to the flux density B of the field; or, expressed differently, χ_m and μ would no longer be constant. An example for a gas (O_2) will be given in Sect. 14.8.

In general, *saturation* of the magnetization M of paramagnetic materials can be observed only at extremely high field strengths and/or very low temperatures. It has been demonstrated for a number of ions, e.g. Cr^{+++}, Fe^{+++}, and Gd^{+++} in sulfate crystals (alums)[1], at $T < 4.0$ K and B up to 5 V s/m^2.

3. *Ferromagnetic materials* are easily recognized even by nonphysicists. They will be attracted by any permanent magnet. Examples: The pure metals Fe, Co, Ni; manganese-containing copper alloys (F.R. HEUSLER, 1898). Physically, the ferromagnetic materials are characterized by the extremely large magnitude of their magnetization M which can be attained. It, in turn, depends in a complex way on the strength of the magnetizing field and the prior treatment of the sample.

We want to demonstrate this dependence experimentally using a *creeping galvanometer* (Sect. 8.4). We use two ring coils as shown in Fig. 14.6; they have the same dimensions and number of turns. The left-hand coil contains an iron core, while the right-hand coil is wound on a wooden core (cf. Sect. 14.1). Both field coils are carrying the same current, and both are surrounded by identical induction loops, but these are connected in opposition. Therefore, the creeping galvanometer indicates the *difference* of the magnetic flux in the two coils, with and without an iron core; i.e. $\Phi_m - \Phi_0$. Division by A, the cross-sectional area of the coils, yields the additional magnetization

[1] W.E. Henry, *Physical Review* **88**, 559 (1952).

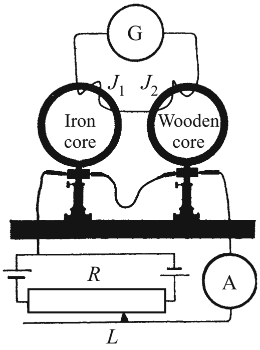


Figure 14.6 Measuring the hysteresis loop of iron using a creeping galvanometer G. Two ring coils as shown schematically in Fig. 14.1. The induction loops J_1 and J_2 are wound oppositely; the galvanometer thus indicates the difference between the two magnetic fluxes Φ, with and without iron. If the slider L is pushed past the midpoint of the resistor, the direction of the current in the two field coils is reversed (why?).

which is due to the iron:

$$M = \frac{1}{\mu_0}(B_\mathrm{m} - B_0)\,. \tag{14.2}$$

We carry out a series of measurements, increasing and then decreasing the field current for both directions, and thus obtain the curve shown in Fig. 14.7, the "hysteresis loop" for wrought iron. From this curve, we read off the following facts:

1. For each value of B_0, the flux density of the empty coil (wooden core), there are two corresponding values of $\mu_0 M$. The right-hand

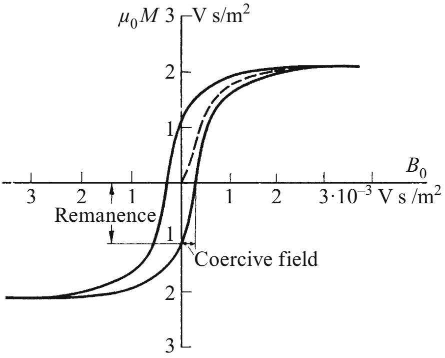


Figure 14.7 A hysteresis loop for wrought iron, measured as in Fig. 14.6. B_0 is the flux density of the empty (i.e. iron-free) field coil. The saturation value of the magnetization (multiplied by μ_0 in the figure) is found at 2.1 V s/m^2. At this value, each iron atom contributes a moment of $m' = M/N_\mathrm{V} = 2.0 \cdot 10^{-23}$ A m^2 (compare Sect. 14.8, Table 14.2). The dashed line indicates schematically a *new curve* or *initial curve* from a sample which had been previously tempered and thus had no remanent magnetization.

branch of the curve in the upper half of the graph (above the B_0 axis) corresponds to increasing magnetization, and the left-hand branch to decreasing magnetization; in the lower half, their roles are reversed (and the magnetization is oppositely directed).

2. The magnetization M approaches a "saturation value" at large values of B_0.

3. A portion of the magnetization M remains when the field of the coil is reduced to zero. It is called the magnetic *remanence* or *remanent magnetization*. The iron core has become a permanent magnet.

4. In order to remove the remanence, the field in the coil must be reversed and its flux density increased to a certain value, called the *coercive field*.

Materials with a small coercive field are termed "magnetically soft". With very pure iron which has been tempered in a hydrogen atmosphere, one can reduce the coercive field to as low as ca. $3 \cdot 10^{-6}$ V s/m^2. In magnetically hard alloys made of Fe, Ni, Co and Al, e.g. Oerstite (a hard magnetic alloy belonging to the AlNiCo family), the coercive field can be increased up to the order of 0.1 V s/m^2, with a remanence of $\approx$ 1 V s/m^2.[C14.7]

C14.7. Modern "hard" materials such as $SmCo_5$ and $Nd_2Fe_{14}B$ attain much higher values of the coercive field and similar values of the magnetic remanence. In addition, their "energy product", a measure of the energy stored in the magnetization (cf. Comment C14.8.), is up to a factor of 10 larger than in the AlNiCo alloys mentioned by POHL. These modern materials are used in many applications today, from small electric motors to the "undulators" employed in free-electron lasers.

5. The cyclic magnetization process, i.e. a full circle around the hysteresis loop, requires that a certain amount of work W be performed. It is equal to the area inside the hysteresis loop, that is[C14.8]

$$W = V \int M \, \mathrm{d}B_0 \qquad (14.6)$$

(V is the volume of the sample).

C14.8. This equation follows from the expression for the energy density in a magnetic field

$$\frac{W}{V} = \int \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{B}\,,$$

whose derivation is given in textbooks on theoretical physics or electrodynamics (e.g. J.D. Jackson, *Classical Electrodynamics* (John Wiley & Sons, NY, 1962), Section 6.2). In vacuum, it leads to the expression for the magnetic field energy given in Eq. (8.28).

6. The old problem of how to levitate a ferromagnetic object freely in a magnetic field was solved only after the discovery of superconductivity (see Comment C14.5.).

An important property of ferromagnetism is its strong temperature dependence. Ferromagnetism vanishes above the "CURIE temperature". This critical temperature lies for example in the HEUSLER alloys below 100 °C. A piece of one of these alloys will stick to

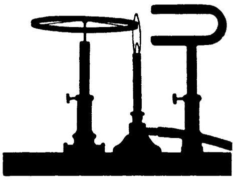


Figure 14.8 A wheel made of nickel and heated on one side will rotate in the field of a permanent magnet. At 356 °C, nickel loses its ferromagnetism, and then the magnet pulls in a cooler, still ferromagnetic region of the wheel. The midplane of the magnet passes through the axle of the wheel; the flame is placed in front of or behind this plane, depending on the desired sense of rotation.

a horseshoe magnet at room temperature, but when it is dipped into boiling water, it falls off. Another demonstration experiment for the critical temperature of ferromagnetism can be seen in Fig. 14.8.

14.5 Definition of the Magnetic Field Quantities *H* and *B* Within Matter. The MAXWELL Equations

The definitions of the magnetic material constants given in Sects. 14.1 and 14.2 are straightforward and correct, but they refer only to the average values of B_0, B_m and H_0 measured in the ring coils used there. We now want to define the vector field quantities $\boldsymbol{H}$ and $\boldsymbol{B}$ in the *interior* of matter, based on experimental results:

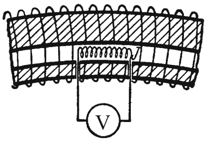
Imagine a ring coil which is initially completely filled with a homogeneous material of permeability μ. We suppose it to contain two small cavities, similar to those already shown in Fig. 13.5 (cf. also the footnote in Sect. 13.5, where you should replace "parallel-plate condenser" by "ring coil" in the present context). One of the cavities is a flat *transverse slit* whose plane is perpendicular to the direction of the field, while the other is a very narrow *axial channel*, parallel to the field direction. Both cavities are again used to accept a measurement probe (real or virtual), indicated by dots in the figure. In both cavities, the fields $\boldsymbol{H}$ and $\boldsymbol{B}$ are measured. In the limit of vanishingly small cavities, these measured fields are defined as the *fields within the matter*. We find that the fields in the two cavities are different! In detail, the results are as follows:

1. The flux density $\boldsymbol{B}$ measured in the transverse slit as a (voltage impulse)/(field area) has the same magnitude and the same direction as the field measured by an induction loop surrounding the coil on its outside, i.e. $B = B_\mathrm{m}$, independently of where the cavity is located within the material. In the limit, we can conclude that the flux density is the same at every point within the material filling the coil. We define it as the *magnetic flux density within matter* and denote it by the vector $\boldsymbol{B}$.

2. On repeating this experiment in the axial channel, we measure a different flux density, one which is not the same as measured with an induction loop outside the coil. However, we also find that the measurement in the axial channel gives a very simple relation for the field $\boldsymbol{H}$: It has the same magnitude and direction as the magnetic field $\boldsymbol{H_0}$ measured in the *empty* ring coil. This is determined as shown in Fig. 14.9. The material filling the coil contains a channel along its long axis which has an arbitrary but constant cross-sectional area[2].

[2] Note for the experimentalist: The ring channel can be cut into the surface of the iron core of the ring coil as an open groove. In practice, this means that one need only make the iron core somewhat smaller in diameter than the inside of the coil. Then the gap between the core and the coil forms the "channel".

Figure 14.9 The equality of the field $\boldsymbol{H}$ in an axial channel to the field in the empty ring coil



The slim induction coil J is placed in this axial channel. It is used to measure, independently of the width of the channel, the same field $\boldsymbol{H}$ as in an empty coil. We can extend this result in a thought experiment to an infinitely thin channel. The field measured in this way is defined as the *field within matter*, and we denote it by the vector $\boldsymbol{H}$. By introducing these vector fields, $\boldsymbol{B}$ and $\boldsymbol{H}$ within matter, we arrive with the aid of Eq. (14.2) at a definition of the vector field $\boldsymbol{M}$, the magnetization:C14.9

C14.9. In textbooks on theoretical physics, it is usual to introduce the magnetization vector $\boldsymbol{M}$ after defining the field $\boldsymbol{B}$. The magnetization is defined in terms of the magnetic moments in the material. Then $\boldsymbol{H}$ is simply an abbreviation for $(\boldsymbol{B}/\mu_0 - \boldsymbol{M})$, which can be used among other things to write the MAXWELL equations in a simpler form.

$$\boldsymbol{M} = \frac{1}{\mu_0}\boldsymbol{B} - \boldsymbol{H}\,. \tag{14.7}$$

This equation demonstrates especially clearly the difference between the fields $\boldsymbol{H}$ and $\boldsymbol{B}$ in matter (see Comment C5.5.). Think for example of the fields within the material of a permanent magnet, for which $\boldsymbol{M}$ is homogeneous and constant (**Exercise 6.3**).

For materials with a constant permeability μ (and thus no hysteresis), we find from Eqns. (14.1) through (14.4) the following relations:

$$\boldsymbol{B} = \mu\mu_0\boldsymbol{H} \tag{14.8}$$

and

$$\boldsymbol{M} = (\mu - 1)\boldsymbol{H} = \chi_{\mathrm{m}}\boldsymbol{H}\,. \tag{14.9}$$

These three equations describe the relationships between the vector fields at every point in space.

With the vector fields $\boldsymbol{H}$ and $\boldsymbol{B}$ defined in the presence of matter, as well as the fields $\boldsymbol{E}$ and $\boldsymbol{D}$ given in Chap. 13, the MAXWELL equations take on their general form:

$$\operatorname{div}\boldsymbol{D} = \varrho\,, \tag{14.10}$$

$$\operatorname{curl}\boldsymbol{E} = -\dot{\boldsymbol{B}}\,, \tag{14.11}$$

$$\operatorname{curl}\boldsymbol{H} = \boldsymbol{j} + \dot{\boldsymbol{D}}\,, \tag{14.12}$$

$$\operatorname{div}\boldsymbol{B} = 0\,. \tag{14.13}$$

Here, ϱ is the charge density of the free charges (as opposed to the bound charges within polarized matter, see Sect. 2.17); and $\boldsymbol{j}$ is the "free" current density (as opposed to the "bound" current density of the molecular currents in magnetized matter, see Sect. 4.4). In empty space, with $\boldsymbol{D} = \varepsilon_0\boldsymbol{E}$ and $\boldsymbol{B} = \mu_0\boldsymbol{H}$, these equations are simplified to those which we set out in Sect. 6.5. The following section demonstrates some of their applications.

14.6 Demagnetization

The vector fields introduced for the interior of magnetic materials will now be investigated in the case that the field coil is *not* completely filled by the material, but instead, only a smaller piece of the material of some arbitrary geometry is placed in the field. In analogy to the geometric properties of the electric polarization, we find that in a magnetic field, also, the observed magnetization depends on the geometry.

An ellipsoid of rotation is brought into a homogeneous magnetic field $\boldsymbol{H}_0$ $(= \boldsymbol{B}_0/\mu_0)$ in such a way that its long axis coincides with the direction of the field vector $\boldsymbol{H}_0$. Then the magnetic field in its interior is homogeneous and we find (in analogy to Eq. (13.13)):

$$\boldsymbol{H} = \frac{1}{1 + N(\mu - 1)} \boldsymbol{H}_0 \, . \tag{14.14}$$

Values for N, here called the *demagnetizing factor*, are collected in Table 13.4. N is a number whose values range from 0 to 1 $(0 \leq N \leq 1)$. Except for the case $N = 0$, the magnetic field $\boldsymbol{H}$ within magnetic materials is thus reduced in comparison to the externally-applied field $\boldsymbol{H}_0$. This is called *demagnetization*.

Some examples may help to elucidate this reduction of the magnetic field $\boldsymbol{H}$ depending on the geometry (as we can see from Eq. (14.14), it is important when $\mu \gg 1$, that is for ferromagnetic substances):

1. For a long, thin rod parallel to the direction of the applied magnetic field $\boldsymbol{H}_0$, $N = 0$ and therefore the field within the material is[C14.10]

$$\boldsymbol{H} = \boldsymbol{H}_0 \, . \tag{14.15}$$

The magnetization

$$\boldsymbol{M}_{\text{rod}} = (\mu - 1)\boldsymbol{H}_0 \tag{14.16}$$

is thus the same as when the material completely fills the field coil, i.e. no demagnetization is observed. $\boldsymbol{B}$ in the rod is equal to $\mu\mu_0\boldsymbol{H}_0$.

2. For a sphere, $N = 1/3$ and thus the magnetic field within the material is given by

$$\boldsymbol{H} = \frac{3}{\mu + 2} \boldsymbol{H}_0 \, , \tag{14.17}$$

and the magnetization is

$$\boldsymbol{M}_{\text{sphere}} = \frac{3(\mu - 1)}{\mu + 2} \boldsymbol{H}_0 = \frac{3}{\mu + 2} \boldsymbol{M}_{\text{rod}} \, . \tag{14.18}$$

Therefore, for $\mu \gg 1$ (ferromagnetic material), the magnetization of a sphere is considerably smaller than that of the rod (**Exercise 14.2**).

[C14.10] This result can also be obtained from Maxwell's equation (14.12). $\boldsymbol{j}$ and $\dot{\boldsymbol{D}}$ are both zero. Then, $\operatorname{curl} \boldsymbol{H} = 0$, and in integral form, $\oint \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{s} = 0$. We choose the closed path to be a long, narrow rectangle, whose long sides are parallel to the long axis of the rod, one of them within the rod and the other outside it. From this, we obtain Eq. (14.15).

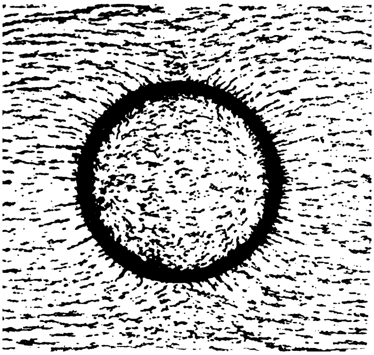


Figure 14.10 Magnetic shielding inside a hollow magnetic shell

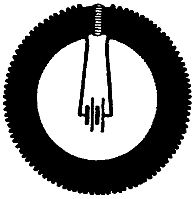


Figure 14.11 Schematic drawing of an electromagnet (ring coil wound on a toroidal core)

3. For a flat disk ($N = 1$), we find[C14.11]

$$\boldsymbol{H} = \frac{1}{\mu}\boldsymbol{H}_0\,, \tag{14.19}$$

and thus for the magnetization:

$$\boldsymbol{M}_{\text{disk}} = \frac{\mu - 1}{\mu}\boldsymbol{H}_0\,; \tag{14.20}$$

then, for $\mu \gg 1$, it is another factor of three smaller than in the sphere.

4. As a practical example of demagnetization, we place a hollow ferromagnetic body, e.g. a hollow iron ball, in a previously homogeneous magnetic field (Fig. 14.10). The field inside the ball vanishes due to the cancellation of the magnetization from the opposite sides, except for a small remnant field. This is the principle of *magnetic shielding*[C14.12] (**Exercise 14.3**).

5. Demagnetization effects also play a role for electromagnets (Fig. 14.11). In the figure, an iron ring of circumference $2\pi r$ is wrapped with N turns of wire which carry a current I. The air gap in the iron core has a width of d, which is sufficiently small that stray fields are negligible. It forms a transverse slit between the two iron poles (Fig. 13.5). The fields B and H within the gap and in the iron core are denoted by the indices d and Fe. We want to compute B_{d}.

From Eq. (14.13), it follows that the magnetic flux density B_{d} is the same as in the filled part of the coil (core)[C14.11],

$$B_{\text{d}} = B_{\text{Fe}}\,. \tag{14.21}$$

C14.11. This can also be found from the MAXWELL equation (14.13), in integral form $\oint \boldsymbol{B} \cdot \text{d}\boldsymbol{A} = 0$. We choose the closed surface to be in the shape of a flat pillbox whose lid and bottom surfaces are parallel to the disk, one within the material and one outside it. It then follows that $\boldsymbol{B} = \boldsymbol{B}_0$, and from this, Eq. (14.19).

C14.12. As an additional example, consider the homogeneous field $\boldsymbol{B}$ within a hollow sphere made of a magnetic material (inside radius a, outside radius b) (for $\mu \gg 1$):

$$\boldsymbol{B} = \frac{9}{2\mu(1 - a^3/b^3)}\boldsymbol{B}_0$$

(see for example J.D. Jackson, *Classical Electrodynamics* (W. de Gruyter, Berlin). 2nd ed., 1983, p. 231). A suitable material would be for example the alloy *supermalloy* (Ni 79, Fe 15.7, Mo 5, Mn 0.3 weight %; its permeability at $B = 2 \cdot 10^{-3}$ V s/m^2 (= T) is $\mu = 10^5$, coercive field $2 \cdot 10^{-7}$ T). A practical application is found in the large field-free chambers used for shielding the earth's magnetic field (and other ambient fields) to allow medical investigations using "SQUID" systems (extremely sensitive superconducting magnetometers).

The field H, however, is discontinuous at the boundaries between the gap and the iron core. We find

$$H_{\mathrm{Fe}} = \frac{1}{\mu} H_{\mathrm{d}} \tag{14.22}$$

(demagnetization).

Both fields, H as well as B, are reduced in comparison to a coil which is completely filled with an iron core (no air gap).[C14.13] We obtain

$$B_{\mathrm{d}} = \mu_0 H_{\mathrm{d}} = \frac{\mu_0 \mu N I}{2\pi r + \mu d}\,. \tag{14.23}$$

Derivation of Eq. (14.23): The relation between the current I and the field H follows from the MAXWELL equation (14.12), where the second term vanishes for a stationary state. In integral form, it is then given by

$$\oint \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{s} = NI\,, \tag{14.24}$$

where the path integral passes around the current NI. In the present case, the integration path follows a circle $2\pi r$ through the iron core and the air gap. We obtain

$$H_{\mathrm{Fe}}(2\pi r - d) + H_{\mathrm{d}} d = NI\,, \tag{14.25}$$

and, together with Eq. (14.22) and the assumption that $d \ll 2\pi r$, we arrive at

$$\frac{H_{\mathrm{d}}}{\mu} 2\pi r + H_{\mathrm{d}} d = NI\,. \tag{14.26}$$

This leads directly to Eq. (14.23).[C14.14]

As the gap width d increases, the field thus decreases. Assuming a constant susceptibility of e.g. $\mu \approx 1000$ for iron, one can find from this equation that even a relatively narrow gap will have a strong effect on the fields, both within the gap and in the iron core itself.

Qualitatively, we have already observed similar behavior on changing the iron magnetic circuit in a coil with an iron core (Fig. 8.12.)

14.7 The Molecular Magnetizability

The different behavior of paramagnetic and diamagnetic materials was already indicated qualitatively in Sect. 14.4. Its quantitative explanation is important among other things for an understanding of *molecular structure* and thus for *chemistry*. For this, we will need the concept of *molecular magnetizability*.

C14.13. In **Video 8.4** http://tiny.cc/xbggoy (see Fig. 8.24), this field reduction can be very clearly seen by making a gap of 0.4 mm width using a few sheets of paper. The force, which is proportional to B^2 (Eq. (8.26)), decreases from more than 530 N to 15 N, i.e. the field is reduced by more than a factor of 6 (**Exercise 14.4**)

C14.14. The depolarization in a filled parallel-plate condenser (Fig. 13.1) can be determined in perfect analogy using the method shown here. We begin with a narrow air gap. At the transition from the air gap to the dielectric, the field $\boldsymbol{D}$ remains unchanged, i.e. it is continuous (Eq. (14.10), see also Comment C14.11), while the field $\boldsymbol{E}$ is reduced due to depolarization. At a constant condenser voltage, the amount of free charges, and thus also $\boldsymbol{D}$, must increase.

In the interior of a body of volume V, an external magnetic flux density $\boldsymbol{B}$, assuming a negligible demagnetization effect (since $\mu \approx 1$) will produce the homogeneous magnetization

$$\boldsymbol{M} = \frac{1}{\mu_0}(\mu - 1)\boldsymbol{B}\,. \tag{14.7}$$

Then the body will acquire a magnetic moment $\boldsymbol{m}$ parallel to the field direction, with

$$\boldsymbol{M} = \frac{\boldsymbol{m}}{V}\,. \tag{14.3}$$

In an atomic picture, the overall magnetic moment $\boldsymbol{m}$ is interpreted as the time-averaged contributions $\boldsymbol{m}'$ of the N individual atoms or molecules, that is

$$\boldsymbol{M} = \frac{N}{V}\boldsymbol{m}' = N_\mathrm{V}\,\boldsymbol{m}'\,. \tag{14.27}$$

Note that $\boldsymbol{m}'$ is a microscopic quantity, while $\boldsymbol{m}$ is the macroscopic magnetic moment of the whole body. We combine the equations (14.7) and (14.27), obtaining

$$\boldsymbol{m}' = \frac{1}{N_\mathrm{V}}\boldsymbol{M} = \frac{(\mu - 1)}{\mu_0 N_\mathrm{V}}\boldsymbol{B}\,. \tag{14.28}$$

Experimentally, we find for diamagnetic materials that μ is constant, and thus the contributions $\boldsymbol{m}'$ are proportional to the magnetic flux density $\boldsymbol{B}$. For this reason, we set

$$\boldsymbol{m}' = \beta\boldsymbol{B} \tag{14.29}$$

and call β the *molecular magnetizability*. We then find

$$\beta = \frac{(\mu - 1)}{\mu_0 N_\mathrm{V}} = \frac{\chi_\mathrm{m}}{\mu_0\,N_\mathrm{V}} = \frac{\chi_\mathrm{m} V}{\mu_0\,N_\mathrm{A} n}\,, \tag{14.30}$$

(for example, β in $\mathrm{A\,m^4/(V\,s)}$, χ_m = magnetic susceptibility, N_V is the number (particle) density of the molecules, V/n the molar volume, N_A the AVOGADRO constant $= 6.022 \cdot 10^{23}\,\mathrm{mol^{-1}}$ (Vol. 1, Sect. 13.1), and $\mu_0 = 1.257 \cdot 10^{-6}\,\mathrm{V\,s/(A\,m)}$).

14.8 The Permanent Magnetic Moments m_p of Paramagnetic Molecules

These can be calculated from the experimentally-determined molecular magnetizability β (Eq. (14.30)). We demonstrate this in the present section.

Without an applied field, the directions of the moments $\boldsymbol{m}_\mathrm{p}$ are randomly distributed in space as a result of thermal motions. The sum of the magnetic moments $\boldsymbol{m}_\mathrm{p}$ averaged over time and space is equal to zero. However, in the presence of a magnetic field, the moments have a preferred direction. As a result, each individual molecule makes a contribution $\boldsymbol{m}'$ averaged over time to the total magnetic moment $\boldsymbol{m}$ of the body. This contribution is usually only a small fraction x of the paramagnetic moment $\boldsymbol{m}_\mathrm{p}$ which every molecule carries.[C14.15] We thus have

$$\boldsymbol{m}' = x\boldsymbol{m}_\mathrm{p}\,. \tag{14.31}$$

This fraction x can be calculated, as long as the interactions between the molecules can be neglected, as is the case in gases and dilute solutions. We then find

$$x \approx \frac{1}{3}\frac{m_\mathrm{p}B}{kT} \tag{14.32}$$

(k = Boltzmann constant $= 1.38 \cdot 10^{-23}$ W s/K, cf. Vol. 1, Sect. 13.10, as well as Sects. 14.6. and 16.6).

The fraction x is thus essentially equal to the ratio of two energies: The work $m_\mathrm{p}B$ is necessary to rotate the carrier of the magnetic moment $\boldsymbol{m}_\mathrm{p}$ to an orientation perpendicular to the field direction. kT is the thermal energy which can be transferred to it by a molecular collision. A precise calculation would have to consider not only a perpendicular orientation, but rather all possible orientations of the molecular moments relative to the field direction (Langevin-Debye formula). This leads in the first approximation to the numerical factor 1/3.

Combining the equations (14.29), (14.31), (14.30), and (14.32) yields

$$\chi_\mathrm{m} = \frac{1}{3}\frac{m_\mathrm{p}^2\mu_0 N_\mathrm{V}}{kT} = \frac{\text{const}}{T}\,, \tag{14.5}$$

that is Curie's law (Sect. 14.4, for N_V = const). In paramagnetic materials, the molecular magnetizability therefore decreases with increasing temperature. Figure 14.12 shows an example (see also Fig. 14.5).

Furthermore, combining Eqns. (14.5) and (14.30) gives an expression for the permanent magnetic moment of a molecule

$$m_\mathrm{p} = \sqrt{\beta \cdot 3 \cdot kT}\,. \tag{14.33}$$

A numerical example for the O_2 molecule: From Table 14.1 (p. 264), we read off the value of the magnetic susceptibility of oxygen referred to the amount of substance, $\chi_\mathrm{m}/(n/V)$, and using Eq. (14.30), we calculate its molecular magnetizability to be

$$\beta = 5.5 \cdot 10^{-26}\,\frac{\mathrm{A\,m^4}}{\mathrm{V\,s}}\,.$$

[C14.15] Note, however, the remarks in small print in Sect. 14.4, Point 2, paramagnetic materials.

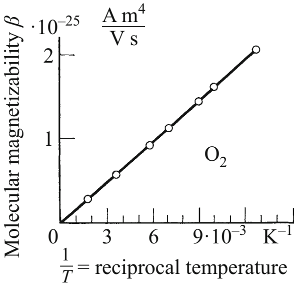


Figure 14.12 The effect of the temperature on the magnetizability of the paramagnetic O_2 molecule (E.C. Stoner, *Magnetism and Matter* (Methuen, London 1934), p. 343. For measurements below 200 K, see E.C. Wiersma *et al.*, *Koninklijke Akademie von Wetenschappen te Amsterdam*, **34**, 494 (1931).)

Table 14.2 Permanent magnetic moments of paramagnetic molecules

Molecule or Ion	NO	O_2	Mn	Fe^{+++}	Ni^{++}	Cr^{+++}
m_p in 10^{-23} A m^2	1.70	2.58	5.40	4.92	3.00	3.54

Inserting this value and room temperature, $T = 293$ K, into Eq. (14.33) gives $m_p = 2.58 \cdot 10^{-23}$ A m^2. Further examples are given in Table 14.2.

> With this value for the moment, we can estimate the fraction x in Eq. (14.32). Compare the analogous computation for electric dipole moments at the end of Sect. 13.10.

14.9 The Elementary Magnetic Moment or *Magneton*. The Gyromagnetic Ratio and the Electronic Spin

In this section, we want to investigate the origins of the permanent magnetic moments discussed above and collected in Table 14.2. We start with the moment which is produced by the orbiting electrons within atoms. If an electron traverses a circular orbit of radius r with the velocity u, then it has an angular momentum (from Sect. 6.6 in Vol. 1) of

$$L = \Theta\omega = mr^2\frac{u}{r} = mru \tag{14.34}$$

(Θ = moment of inertia, ω = angular velocity, m = mass).

In addition, according to Eq. (4.4), it gives rise to a ring current of $I = -eu/l = -eu/2\pi r$ ($e = 1.602 \cdot 10^{-19}$ A s is the elementary charge, Sect. 3.6), and this current is associated with a magnetic moment of

$$m_p = IA = -\frac{eu}{2\pi r}\pi r^2 = -\frac{1}{2}eur\,. \tag{14.35}$$

The quotient

$$\frac{m_\mathrm{p}}{L} = \frac{\text{Magnetic moment of the particle}}{\text{Angular momentum of the particle}} \qquad (14.36)$$

is called the *gyromagnetic ratio*. For an electron in a *circular orbit*, combining Eqns. (14.35) and (14.34) gives for the gyromagnetic ratio

$$\frac{m_\mathrm{p}}{L} = -\frac{1}{2}\frac{e}{m} = -8.8 \cdot 10^{10}\,\frac{\mathrm{A\,s}}{\mathrm{kg}} \qquad (14.37)$$

(e/m, the specific charge of the electron, is equal to $1.76 \cdot 10^{11}$ A s/kg).

In BOHR's model for the hydrogen atom (e.g. as described by POHL in the 13th edition of "*Optik und Atomphysik*", Chap. 14), an electron on the innermost circular orbit in the hydrogen atom has the elementary angular momentum

$$L = \frac{h}{2\pi} \qquad (14.38)$$

($h =$ PLANCK's constant, $6.626 \cdot 10^{-34}$ W s^2 $= 4.36 \cdot 10^{-15}$ eV s).

Substituting this expression into Eq. (14.37) yields the *elementary magnetic moment* or BOHR *magneton*

$$m_\mathrm{Bohr} = \frac{e}{m}\frac{h}{4\pi} = 9.27 \cdot 10^{-24}\,\mathrm{A\,m^2}\,. \qquad (14.39)$$

The measured values in Table 14.2 are of the order of magnitude of the BOHR magneton. We consider this agreement to be an indication that the orbital angular momentum can play an important role in forming the magnetic moments of atoms and molecules. An additional contribution is made by the electron's *spin*, which we will discuss in the following.

The *measurement* of a gyromagnetic ratio was first carried out for a *ferromagnetic* substance, iron. With iron, we observed in Sect. 4.4 the first of the phenomena which are now termed "gyromagnetic": A *ferromagnetic object during the process of magnetization acquires not only a magnetic moment, but simultaneously a mechanical angular momentum*. They are the sums of all the magnetic moments m_p and all the angular momenta L of the N participating electrons in the object.C14.16

C14.16. See the EINSTEIN-de HAAS effect (Sect. 4.4). The inverse effect has also been observed: A rotating iron rod becomes magnetized (BARNETT effect).

For the quantitative evaluation, we proceed as follows: The rod in Sect. 4.4 has the moment of inertia Θ. At the remanent magnetization, the rod acquires an angular momentum of $NL = \Theta\omega_0$. Here, NL is the angular momentum of all N participating electrons. The rod leaves its rest position with the maximum value ω_0 of its angular velocity and reaches its impulse deflection of α_0. The quantity ω_0 is found from the relation $\omega_0 = \omega\alpha_0$, where ω is the circular frequency of the rod which acts as a torsional pendulum. After measuring the angular momentum NL, we take the rod out of the field coil and measure its remanent magnetic moment $m = Nm_\mathrm{p}$, e.g. by using the procedure described Fig. 8.15. (In practice, it is expedient to use an alternating current in the coil, causing a periodic force to act on the magnetic moments in the rod. The rod (torsional pendulum) thus responds with forced oscillations (Vol. 1, Sect. 11.10). Adjustment of the frequency of the current in the coil to the resonance frequency of the torsional pendulum gives a large increase in sensitivity to the rotations of the rod (*resonance enhancement*)).

In this way, we obtain experimentally the gyromagnetic ratio of an electron in iron:

$$\frac{m_\mathrm{p}}{L} = -1.75 \cdot 10^{11}\,\frac{\mathrm{A\,s}}{\mathrm{kg}} = -\frac{e}{m}\,. \tag{14.40}$$

Today, it has proven to be expedient to quote only *relative values* of gyromagnetic ratios: A measured value is referred to the gyromagnetic ratio of an electron in a *circular orbit* and defines its "g-factor" (often called the Landé g-factor) by the equation

$$g = \frac{m_\mathrm{p}}{L} \Big/ \left(\frac{m_\mathrm{p}}{L}\right)_{\text{circular orbit}}\,. \tag{14.41}$$

We find the value $g = 2$ for the electrons in iron using Eqns. (14.40) and (14.37). (Precision measurements on free electrons later gave the value $g = 2.0023$.)

The gyromagnetic ratio of an electron in ferromagnetic iron is thus practically twice as large as would be expected from the orbital current of the electron on a Bohr orbit. This gyromagnetic ratio therefore cannot originate from the *orbital motion* of the electrons, as in Bohr's model. Instead, the electrons must have an angular momentum L and an associated magnetic moment m_p even when they are not moving on circular orbits, i.e. when their centers of gravity are at rest. Both quantities are most simply explained as the result of a *proper rotation*, like that of a spinning top. For that reason, the angular momentum of a non-orbiting electron has been given the name *spin*.C14.17 Its magnitude was determined by making use of the *directional quantization* found experimentally by W. Gerlach and O. Stern (see e.g. evunix.uevora.pt/~stadler/FAN-06-07/History_of_the_Stern-Gerlach_experiment.html). The *spin of an electron* is

C14.17. To distinguish between these two angular momenta (of the orbital motion and the proper rotation), the terms *orbital angular momentum* and *proper angular momentum* or *spin* have been introduced.

$$L_\mathrm{S} = \frac{1}{2}\frac{h}{2\pi}\,. \tag{14.42}$$

Inserting Eq. (14.42) into Eq. (14.40) yields the *permanent magnetic moment of the electron associated with its spin*:

$$m_\mathrm{p} = -\frac{h}{4\pi} \cdot \frac{e}{m}\,. \tag{14.43}$$

It thus has the same magnitude as the elementary magnetic moment named for Bohr (Eq. (14.39)), m_Bohr (often denoted as μ_B).

14.10 The Atomic Interpretation of Diamagnetic Polarization. LARMOR Precession

Now, how does diamagnetism come about? In diamagnetic atoms, the electrons in the electronic shells have their permanent spin moments pairwise oppositely directed (antiparallel); furthermore, their orbital moments cancel pairwise. Only in this way is it possible that the electronic shells have no overall permanent magnetic moments. A moment is formed by *induction* when the atom is brought into a magnetic field.

In Fig. 14.13, the equatorial plane of a planar atomic model is shown as a shaded disk. The symmetry axis A of the atom is tilted by an arbitrary angle ϑ relative to the direction of the field $\boldsymbol{B}$. At the moment represented by the drawing, an electron is at a distance r_n from the vector $\boldsymbol{B}$. For simplicity, we assume that the magnetic field increases linearly with time after being switched on; its maximum strength B is attained after a time Δt. During the period when the field is increasing, Eq. (6.2) applies. There is thus an electric field along the circumference of the circle $2\pi r_\mathrm{n}$ with the magnitude

$$E = \frac{\dot{B}\pi r_\mathrm{n}^2}{2\pi r_\mathrm{n}} = \frac{r_\mathrm{n}}{2}\dot{B}\,. \tag{14.44}$$

It causes the electron to accelerate:

$$a = \frac{Ee}{m} = \frac{1}{2}\frac{e}{m}\cdot r_\mathrm{n}\dot{B}\,. \tag{14.45}$$

This acceleration increases its orbital velocity along the circular orbit $2\pi r_\mathrm{n}$ within the time Δt by an amount

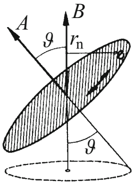
$$u = \frac{1}{2}\frac{e}{m}r_\mathrm{n}B\,, \tag{14.46}$$

corresponding to the angular velocity $\omega = u/r_\mathrm{n}$, that is

$$\omega_\mathrm{Larmor} = \frac{1}{2}\frac{e}{m}B\,. \tag{14.47}$$

This angular velocity or circular frequency, named for its discoverer, is independent of the radius r_n; it holds for all the elementary charges

Figure 14.13 The origin of LARMOR precession. The double arrow indicates that the electrons in a diamagnetic molecule are circling as pairs in opposite directions.



within the atom. As a result, all those elementary charges rotate together, that is the atom as a whole rotates around the axis defined by the magnetic field. An observer supposed to be at the center of the atom would see no changes in the electronic orbits. The orbits of the electrons within an atom can have a common axis of rotation. If this axis is not parallel to the direction of $\boldsymbol{B}$, then it moves around that direction $\boldsymbol{B}$ on a precession cone. The sense of the rotation ensures that the magnetic moment that it generates opposes its own origin, the magnetic field (LENZ's law, Sect. 8.3). According to equations (14.9) and (14.7), this means that χ_m is negative and $\mu < 1$, as observed for diamagnetic materials.

We can thus explain both paramagnetism and diamagnetism with simple models.

14.11 Ferromagnetism, Antiferromagnetism, and Ferrimagnetism

When paramagnetic atoms are bound together in a solid, they as a rule continue to be paramagnetic. As an example, we mention the triply-ionized Gd ion, Gd^{3+}, in the alum $\mathrm{Gd_2(SO_4)_3{\cdot}8H_2O}$[3] Here m', the average contribution of an ion to the overall magnetic moment m, which is given by $m' = M/N_\mathrm{V}$ (Eq. (14.27)), increases in an applied magnetic field proportionally to B (Fig. 14.14, top) and inversely proportionally to the temperature T (CURIE's law, Fig. 14.15A′). Only at low temperatures and in high magnetic fields does m' attain the order of magnitude of a BOHR magneton, m_Bohr (Fig. 14.15A).

Ferromagnetic solids (Sect. 14.4) exhibit a behavior which is surprisingly different. As an example, we choose *nickel*. At a small

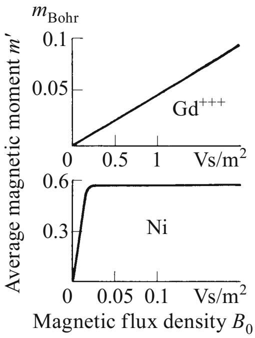


Figure 14.14 The average contribution m' of a Ni atom and a Gd^{3+} ion to the overall magnetic moment $m = MV$ (Eq. 14.9). ($m' = M/N_\mathrm{V}$, in units of the BOHR magneton m_Bohr; $T = 300\,\mathrm{K}$). *Top*: $\mathrm{Gd_2(SO_4)_3 \cdot 8H_2O}$ (measured as described in Sect. 14.3); *bottom*: Ferromagnetic, microcrystalline nickel (magnetic flux density B_0 defined as in Fig. 14.7 ($B_0 = \mu_0 H$)).

[3] Crystalline gadolinium sulfate octahydrate (see also Sect. 14.4, Point 3.2). Its density is $\varrho = 3.01\ \mathrm{g/cm^3}$ and the particle number density of the Gd ions is $N_\mathrm{V} = 4.86 \cdot 10^{21}\ \mathrm{cm^{-3}}$.

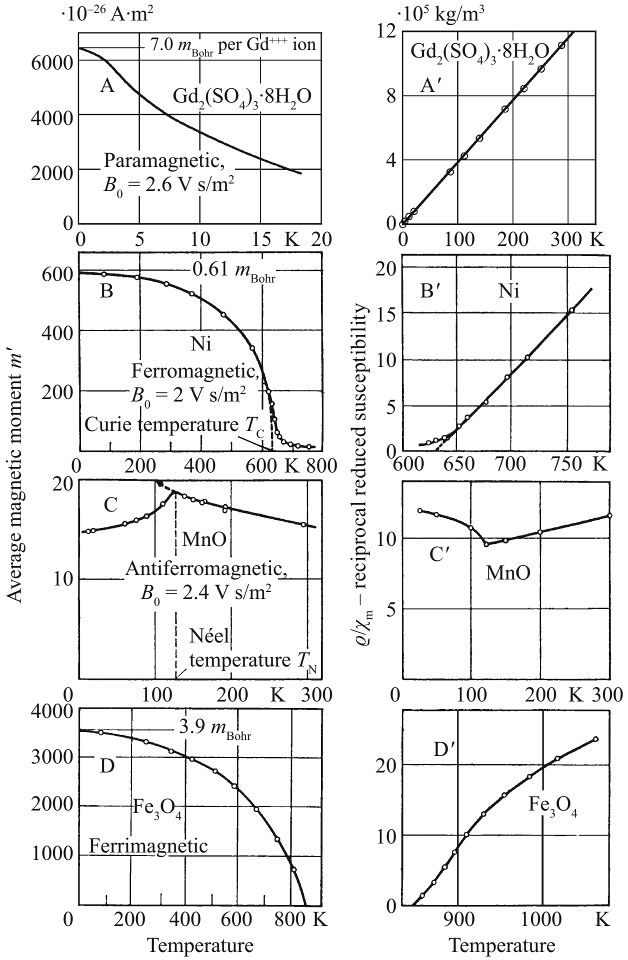


Figure 14.15 Influence of the temperature on solids with different magnetic behaviors: The average contribution m' of one molecule or one atom or ion to the overall magnetic moment m, and the reciprocal magnetic susceptibility (ϱ/χ_m) relative to the density of the material (the flux density B_0 is defined as in Fig. 14.7). (Literature to Parts A and A′: see *Landolt/Börnstein*, 6th edition, Vol. II/9, "*Magnetische Eigenschaften*" (Springer-Verlag 1962), in the article by I. Grohmann and S. Hüfner, pp. 3-200 ff. For the other parts of the figure, see E. Kneller, "*Ferromagnetismus*" (Springer-Verlag 1962), Chap. 4–6).

fraction of the magnetic flux density which we required for Gd^{3+}, the average atomic contribution m' in Ni has already reached its saturation value (Fig. 14.14, bottom). It increases below the CURIE temperature ($T_\mathrm{C} = 631\,\mathrm{K}$) with decreasing temperature, and already at room temperature, it has a value of $0.58\,m_\mathrm{Bohr}$ (Fig. 14.15, B). A similar behavior is observed for iron: In Fig. 14.7, already at

$B_0 = 3 \cdot 10^{-3}$ V s/m^2 and room temperature, $m' = 2.0 \cdot 10^{-23}$ A m^2 = 2.1 m_{Bohr} is observed.

When the average contributions m' from the individual atoms to the overall magnetic moment m reach the order of magnitude of m_{Bohr}, then a large fraction of the atomic moments must already be oriented parallel to each other. Only very large magnetic fields at very low temperatures can achieve such a parallel orientation; that is shown by materials which behave paramagnetically. There remains only one explanation: *In solids with ferromagnetic behavior, new forces appear within the crystal lattice. They bring about an orientation of the atomic magnetic moments within microscopic regions of the crystal. These regions (*WEISS *domains or magnetic domains) are spontaneously magnetized up to saturation.* Macroscopically, the directions of the magnetization M of all the domains are distributed statistically over all angles; therefore, a ferromagnetic crystal is "as a whole" not magnetized before it is brought into an external magnetic field[4]. Thermal motions limit the spontaneous magnetization, i.e. the formation of domains in which all the microscopic elementary magnetic moments are oriented parallel to each other, and thus for example there is not some fraction of them which is antiparallel to the rest. An external field can now align the directions of spontaneous magnetization, which is present *in spite of the thermal motions*, so that all the domains are oriented in the same direction, parallel to the field. The saturation value of the magnetization M which can be observed in an applied field decreases from its maximum value at $T = 0$ K monotonically up to the CURIE temperature T_C (Fig. 14.15B). Above this CURIE temperature, the crystal behaves simply as a paramagnet (Fig. 14.15B′).[C14.18]

C14.18. In this temperature range, the magnetic susceptibility χ_m of the ferromagnet is described by the CURIE-WEISS law: $\chi_m = C_T/(T - T_C)$;
C_T is the CURIE constant, and T_C is the CURIE *temperature*.
A recent review of magnetic microstructures, as well as an overview of ferromagnetism, can be found in the book "Magnetic Domains" by Alex Hubert and Rudolf Schaefer, Springer-Verlag (1998).

The assumption of microscopically small, magnetically saturated crystal regions in materials which exhibit ferromagnetic behavior is not new (I.A. EWING, 1891).[C14.18] Two points have emerged in more recent times: 1. The recognition that the alignment of the elementary magnetic moments within the spontaneously saturated regions cannot be explained by magnetostatics alone; and 2. The experimental means to observe these saturated regions and the direction of their spontaneous magnetization microscopically. These two points will be discussed in the following.

In order to visualize microscopically the spontaneously-magnetized crystal regions (domains), one polishes the surface of a non-magnetized iron crystal, preferably using electropolishing. Then one brings a suspension of extremely fine ferromagnetic Fe_2O_3 power onto the surface. The powder achieves the same result for micro-

[4] This holds of course only for objects which contain many spontaneously-magnetized domains. If the body is in the form of fine powder in which each powder particle contains only one spontaneously-magnetized domain ("single-domain particles"), then the particles act like giant paramagnetic molecules with very large magnetic moments m_p; this is called *superparamagnetism*. The smaller the particles, the lower their CURIE temperature, since not the thermal motions, but instead the surface tension limits their spontaneous magnetization.

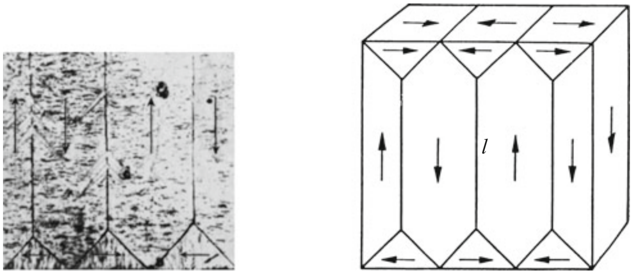


Figure 14.16 *Left*: Visualization of the domain boundaries between spontaneously-magnetized regions, and the directions of the magnetization *M* within them. *Right*: An explanatory sketch. In polycrystalline material, the domain boundaries often assume rather complex shapes.

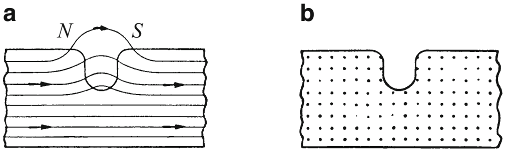


Figure 14.17 Furrows, scratches etc. produce *poles N S* on the surface of a magnetized object (so long as they are not parallel to the direction of the magnetization, which is perpendicular to the plane of the page in Part b)

scopic dimensions as iron filings on a macroscopic scale (Chap. 4). Figure 14.16 shows an image of this type on the left, and on the right, a sketch which explains the image on the left. At the domain boundaries, poles occur. The dark powder collects within their magnetic fields.

These images not only allow us to recognize the *boundaries* of the domains, but also the *directions* of the magnetization *M* within them. For that purpose, a brush with glass bristles was used to scratch fine furrows on the surface and thereby to localize magnetic poles along them. Such poles however are formed only when the furrows are nearly perpendicular to the direction of the local magnetization *M*, as can be seen from Fig 14.17a. Therefore, the darker lines filled with the powder can be seen in Fig. 14.16 only where the furrows cross the magnetization directions (arrows in the sketch).

In order to investigate the magnetization process, let us first heat a ferromagnetic object (made e.g. of iron) temporarily to above its CURIE temperature; after cooling, it will be in an overall non-magnetized state. In which manner can an applied magnetic field then convert this macroscopically non-magnetized object into a magnet, so that it acquires an overall magnetic moment *m*?

The microscopically observed, spontaneously-magnetized domains in Fig. 14.16 (left) are not bounded by surfaces in the mathematical sense, but rather by separation layers of finite thickness (ca.

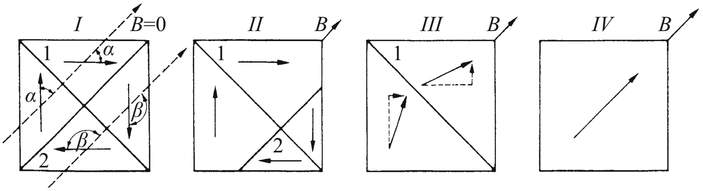


Figure 14.18 The magnetization process. The lines denoted by 1 and 2 indicate domain walls. The arrows show the direction and relative strength of the magnetization M. The short arrows in Part III indicate the sense of rotation.

0.1 μm), called *domain walls* or BLOCH walls. The transition of the direction of magnetization of a domain into that of its neighboring domain occurs within these domain walls. The intersections of the walls with the surface of the sample show relatively simple patterns in Fig. 14.16, since the surface was cut parallel to one of the cubic crystallographic faces. The sketch at the right in Fig. 14.16 which explains the observed pattern can be reduced to a schematic in which the lengths l have become zero; this schematic can be seen in Fig. 14.18.

Now suppose that the iron cube is placed in a magnetic field $\boldsymbol{B}$ which is parallel to a face diagonal of the cube. This field gives the macroscopic magnetization of the cube a preferred direction and thus produces an overall magnetic moment m which can make its effects felt in the space around the cube. How does this occur?

Of the two angles between the field $\boldsymbol{B}$ and the local magnetization vectors $\boldsymbol{M}$, α is smaller than β. As a result, the regions containing α grow at the cost of those containing β; this occurs through a shift of the domain wall 2 to the right (in the figure, Part II). In Part III, only two domains are still present. Their magnetization vectors $\boldsymbol{M}$ initially have the same orientations as those on the left and above in Part II; they are *symmetric* to the remaining domain wall 1. Therefore, this wall will not be shifted further; instead of a *wall motion*, a new process occurs: The directions of magnetization *rotate* as the field increases further, and both domains approach the direction of the field (Part III). This *domain rotation process* comes to an end when the whole crystal has a unified magnetization (Part IV), i.e. its magnetization is saturated (**Exercise 14.5**).

The wall motion during the process of magnetization can be observed and recorded microscopically. An impressive effect is the “sticking” of domain walls on disturbances in the crystal lattice, such as nonferromagnetic inclusions. When the walls break loose irreversibly, there are sudden, jerky jumps in the magnetization $\boldsymbol{M}$. These also can be readily observed.

A polycrystalline material, a thin, soft iron wire Fe, is shown in Fig. 14.19, surrounded by an induction coil J. The coil is connected to an amplifier and thence to an oscilloscope with its horizontal deflection proportional to the time, and also to a loudspeaker. N-S is a small bar magnet; it can be moved back and forth in the direc-

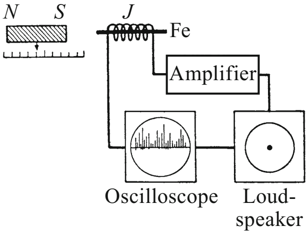


Figure 14.19 The demonstration of statistically distributed jumps in the magnetization during a uniform variation of the magnetizing field (BARKHAUSEN effect)

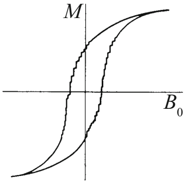


Figure 14.20 Resolution of the BARKHAUSEN effect (steps in the magnetization) in a hysteresis loop, schematic

tion towards or away from the iron wire, along a ruled scale. As it approaches the iron wire, the magnetization of the wire increases. When the magnet is removed, the magnetization again decreases. In this way, the hysteresis loop of the iron wire can be repeatedly traversed. Earlier, in Fig. 14.7, the loop appeared to ba a smooth curve. Now, we find that it is composed of many small steps, as can be seen in Fig. 14.20; each step represents an irreversible jump due to a jerky wall motion. These motions can be seen on the oscilloscope as individual steps along the curve. In the loudspeaker, they produce a crackling noise which is named for H. BARKHAUSEN (the Barkhausen effect). Most of these jumps occur in a purely statistical manner. A few of the strongest, however, occur again and again at the same positions of the bar magnet.

Domain-wall motions occur predominantly in weak applied magnetic fields, while domain rotations occur in stronger fields. There, the situation is complex. Briefly, one can summarize the whole process as a *magnetic recrystallization*, which usually begins at locally-formed "nucleation centers" (they play a role not only in crystallization, but also in evaporation and condensation processes).

Based on these experimental findings, we can readily understand the sketch in Fig. 14.21A: Iron crystallizes in the body-centered cubic (bcc) system, i. e. its crystal lattice consists of two parallel, intermeshing cubic sublattices. The vertices of the cubes of the second lattice lie at the intersections of the body diagonals of the first lattice. The cubes sketched in the figure belong to a magnetic domain which is spontaneously magnetized to saturation. Instead of the atoms, only the magnetic moments of the atoms are drawn at the vertices of the cubes, differently colored and readily distinguishable for the two sublattices.

In Fig. 14.21, Parts B and C complement Part A: They contain the magnetic structures which are discussed in the following. In Fig. 14.21B, equally strong atomic moments in the two sublattices

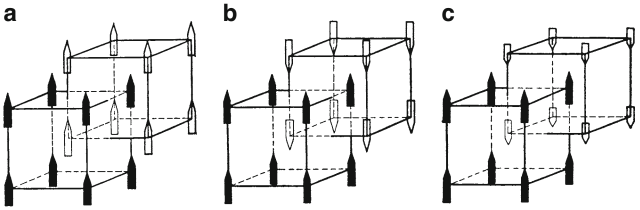


Figure 14.21 Ferromagnetic (A) and antiferromagnetic (B) ordering of the atomic magnetic moments in a body-centered cubic crystal lattice; C: Schematic representation of a ferrimagnetic material

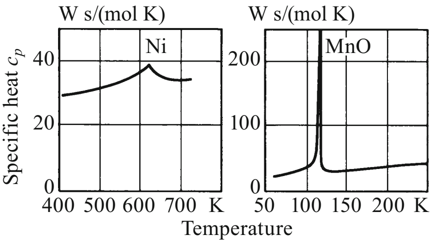
are oriented antiparallel to each other. Each of the two sublattices is spontaneously magnetically saturated, but the resultant overall magnetic moment of the domain is zero. This is the simplest case of a body which exhibits *antiferromagnetic* behavior. The antiferromagnetic order is reduced by random thermal motions, just like ferromagnetic order, as the temperature increases. The long-range order vanishes at a temperature T_N, the *Néel temperature* (named for Louis NÉEL). Above T_N, antiferromagnetic materials behave as paramagnets in an applied magnetic field.C14.19

C14.19. In antiferromagnets, the magnetic susceptibility χ_m in the temperature range above T_N is described by $\chi_\mathrm{m} = C_T/(T + \Theta)$, where Θ, the CURIE-WEISS temperature, is positive.

Also at temperatures below T_N, an antiferromagnet behaves at a constant temperature like a *paramagnet*, i.e. its magnetization $\boldsymbol{M}$ increases linearly with the applied flux density $\boldsymbol{B}$. However, the influence of the temperature on M/N_V and on the reduced susceptibility χ_m/ϱ is more complex than for a simple paramagnetic material. Figures 14.15C and C′ give examples for antiferromagnetic MnO.

With decreasing temperature below the NÉEL temperature, $1/\chi_\mathrm{m}$ again increases (Fig. 14.15C′). The reason: The orientation of the atomic magnetic moments m_p in an applied magnetic field becomes less effective when the antiferromagnetic coupling between the moments m_p of the atoms increases relative to the thermal energy.
The specific heat capacity of the solids with ferromagnetic and antiferromagnetic behavior is anomalously large in the neighborhood of the CURIE or the NÉEL temperature (Fig. 14.22). Explanation: In order to destroy the ordering of the atomic magnetic moments, heat must be added to the sample.

Figure 14.22 Variation of the specific heats of ferromagnetic and antiferromagnetic materials near their magnetic transition temperatures (the CURIE temperature and the NÉEL temperature, respectively)



There are solids in which the two sublattices are not equivalent, but instead have differing overall resultant magnetic moments (Fig. 14.21C). Then the difference between the sublattice magnetizations gives rise to a spontaneous magnetization. Such materials are termed *ferrimagnetic* (*ferrites*). A well-known example is *magnetite*, known for over 2000 years as "lodestone", Fe_3O_4.

This mineral has the spinel structure. Its chemical formula is $FeO{\cdot}Fe_2O_3$. The negative oxygen ions form a face-centered cubic (fcc) sublattice, into which one divalent and two trivalent iron ions per formula unit are included. The divalent iron ions can be partially or completely replaced by other metal ions; this gives rise to a large variety of cubic ferrites, as these compounds are called.
The ferromagnetism of the ferrites comes about as follows: One half of the trivalent iron ions forms one sublattice, while the other half forms another sublattice together with the divalent metal ions. The divalent oxygen ions are diamagnetic, and thus have no permanent magnetic moments m_p. The magnetic moments on the metal ions are oriented antiparallel to each other in the two sublattices. The moments of the trivalent iron ions thus mutually cancel. The resulting spontaneous magnetization is formed by the magnetic moments of the divalent metal ions.

In ferromagnetic materials, the magnetization $\boldsymbol{M}$ can be increased only up to a maximum or saturation value by applying an external magnetic field of increasing flux density $\boldsymbol{B}$. The same is true of ferrimagnetic materials. The temperature dependence of the saturation values of $\boldsymbol{M}$ is shown in the curve in Fig. 14.15D; it is similar to that for ferromagnetic materials. However, the influence of the temperature on the reduced susceptibility χ_m/ϱ (Fig. 14.15D′) is different from the case of ferromagnets (Fig. 14.15B′).

As a result of the inequivalence of the two antiferromagnetically coupled sublattices, not only are the saturation values of their magnetic moments m different, but also so is the temperature dependence of the spontaneous magnetization of the sublattices. For this reason, it can happen that the resulting spontaneous magnetization goes to zero with increasing temperature *before* the CURIE temperature is reached. This can be shown using a lithium-chromium ferrite in a surprising demonstration experiment.

In Fig. 14.23, a rod made of this material is hung from a thread. The rod has a remanent moment m. In field-free space, it will be oriented perpendicular to the plane of the page. Between two magnetic poles, it orients itself parallel to the plane of the page, for example with the arrow pointing to the right. Then the rod is warmed by hot air and thermal radiation from a glowing heating coil beneath it. At $T = 38\,°\mathrm{C}$, it turns until it is again perpendicular to the plane of the page; it has thus become non-magnetic. If the temperature is increased still further, the rod again turns into the plane of the page, but this time with the arrow pointing to the left. Therefore, the overall magnetic moment m of the rod has reversed its direction by 180°.

The ferrites are ceramic-oxide materials (first described in 1779). As semiconductors, they have resistivities which are many orders of magnitude greater than those of metals. Therefore, perturbations caused by eddy currents play no role in their properties.

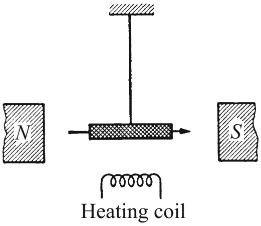


Figure 14.23 The magnetic moment of a ferrite is formed as the difference between two oppositely-directed magnetic moments with differing temperature dependencies. At $T > 38\,°C$, the ferrite rod rotates by 180° (here, the ferrite is $Li_2^+ \cdot Cr_6^{2+} \cdot Fe_6^{3+} \cdot O_{16}^{2-}$).

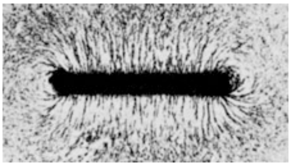


Figure 14.24 A "compass needle" whose long axis points from east to west, because its north and south poles are located on its sides (**Exercise 14.6**)

This makes the ferrites exceptionally important as materials for high-frequency and communications technology. Complex ferrites, e.g. $PbO \cdot 4Fe_2O_3 \cdot BaO \cdot 6Fe_2O_3$, can be powdered and bound together with a glue, and they have very high coercive fields. They are widely applied for example in magnetic data storage or as low-cost materials for making strong permanent magnets.

For example, one can fabricate short, thick permanent magnets from ferrites with a high coercive field, although this shape has a large demagnetizing factor (Sect. 14.6). An example is shown in Fig. 14.24.

Exercises

14.1 In order to determine the magnetic susceptibility χ_m of graphite (arc-lamp carbon), the weight of a sphere made of this material with a diameter of $d = 6\,\text{mm}$ is measured in an inhomogeneous magnetic field (Fig. 14.3). For the magnetic flux density B at the position of the sphere, we find $B = 2.0\,\text{T}$, and for the field gradient, $dB/dz = -20.0\,\text{V s/m}^3$ (the z direction is oriented upwards, opposite to the force of gravity). When the magnetic field is switched on, the weight of the sphere decreases by $5.1 \cdot 10^{-5}\,\text{N}$ ($\approx 5\,\text{mg} \cdot g$, i.e. $\approx 2\,\%$). Calculate χ_m. Is it necessary to take the demagnetizing factor into account? (Sects. 14.3, 14.6)

14.2 a) In analogy to the derivation of the electric field E_p in the interior of a uniformly polarized sphere, as described in Sect. 13.6 (Eqns. (13.12) through (13.16)), derive the magnetic field H_m in the interior of a uniformly-magnetized sphere (a permanent magnet with magnetization M), beginning with Eqns. (14.14) to (14.17).
b) What do you find in this case for the flux density B_m within the sphere? (Sect. 14.6)

14.3 As a simple example of magnetic shielding against an external magnetic field H_a, the field H_i within a spherical cavity in a magnetizable plate of permeability μ_a is to be computed. The plate is oriented perpendicular to H_a (analogously to Fig. 13.8). H_a is parallel to H_0, as in the electrical case; Eq. (13.18) holds also for magnetic fields (see e.g. M.H. Naifeh and M.C. Brussel, *Electricity and Magnetism* (John Wiley, New York 1985), p. 308). Find the relationship between H_i and H_a. (Sect. 14.6)

14.4 With the pot magnet in Fig. 8.24 (Video 8.4), we want to estimate the forces F_0 for a gap width of zero and F_d for a gap width $d = 0.4$ mm. The permeability of the iron is $\mu = 727$, the number of turns in the magnet coil is $N = 175$, and the current is $I = 0.75$ A. The end surface of the inner pole has an area of $A = 7.55\,\text{cm}^2$ and is the same as that or the outer pole (the ring), so that Eq. (14.23) may be applied. The length of the path integral (Eq. (14.26)) is $l = 12$ cm (note that in the calculation of the path integral, the quantity d occurs twice). (Sect. 14.6)

14.5 In Video 10.1, "The inertia of the magnetic field", the slow buildup and decay of a magnetic field are shown (Sect. 10.2, Comment C10.5).

a) Initially, switch 1 (Fig. 10.6) is closed and the increase of the current up to its maximum value of 15 mA is followed (the OHMic resistance of the coil is $R = 130\,\Omega$). Evaluate this rise by plotting it on semilogarithmic graph paper and test it for an exponential dependence.

b) In a second experiment, the decay of the current is determined by closing switch 2 and shortly thereafter opening switch 1. Again, look for an exponential dependence.

c) Finally, the leads to the coil are exchanged, switch 2 is opened and switch 1 closed, and again the rise in the current is registered. What time dependence is found now? How can you explain qualitatively the observed time dependence in all three experiments? (Sect. 14.11)

14.6 Explain why the magnetic flux density B in front of the ceramic disk in Fig. 14.24 is smaller than that in front of a long bar magnet with the same homogeneous magnetization M. As a simplification, approximate the shapes of the disk and the bar by ellipsoids of rotation. For the disk, the ratio of thickness to diameter is $l/d = 0.1$. (Sect. 14.11)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_14) contains supplementary material, which is available to authorized users.

Optics

II

Introduction. Measuring the Optical Radiant Power 15

15.1 Introduction

Some black night in your darkened bedroom, stick your head under the covers and press on the corner of your eye. Then you will *see* a *bright light*, in the form of a *colored, yellow, shining* ring. The words printed here in italics are used by our natural language to describe *sensory perceptions*. Every involvement with *light* and its measurement (photometry, see Chap. 29), and every investigation of *colors* and *brilliance* do not belong in the realm of physics; they are the jurisdiction of psychology and physiology. Taking this fundamental fact into account from the outset can avoid many fruitless discussions and diversions.

The usual excitation of our visual perceptions, *light, brightness, color and brilliance* is caused by a form of *radiation*. Originating with radiating objects or light sources, "something" arrives at our eyes. It requires no kind of transporting medium on its way there. The radiation of the sun and the stars reaches us through the vast emptiness of space. Today, school children learn that this radiation consists of electromagnetic waves of very short wavelengths. We often call this *light*-producing radiation "optical radiation", or simply light. The word *light* in the sense of radiation is used even for invisible rays (ultraviolet, infrared). This double meaning, *light* as a sensory perception, and light as a physical radiation, corresponds to the language convention in acoustics, as well (Vol. 1, Sects. 12.24–12.30). There, also, the perception of *sound* is excited by a type of radiation. The radiation which causes the perception of sound is usually called "sound waves", acoustic radiation, or simply sound. In this case, also, the word *sound* is used without hesitation for inaudible sounds (ultrasound, infrasound).

15.2 The Eye as a Radiation Detector. MACH's Stripes

Our eyes can accomplish a great deal in the physical investigation of the radiation which excites our sensory perception of *light*. They can take us much further than our ears do in the analogous problems of

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_15

Figure 15.1 The origin of MACH's stripes. When the disk is rapidly rotated, the image which is shown in Fig. 15.2 as a photograph is seen.

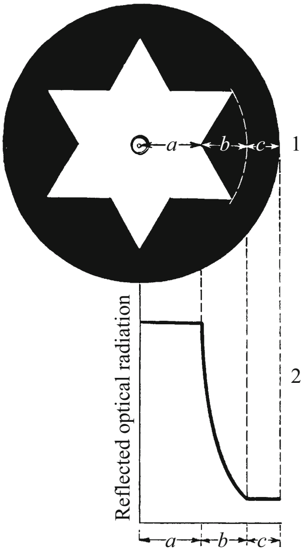
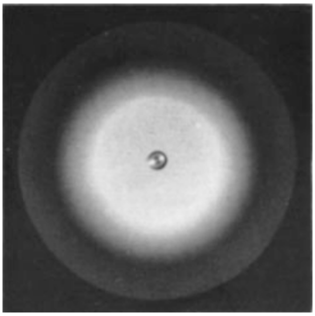


Figure 15.2 MACH's stripes at the boundaries between white and grey, and between grey and black



acoustic radiation. But, like every organ of sensory perception, our eyes fail when it comes to quantitative measurements. They fail to give us an exact numerical description of "less" or "more".

A drastic example is provided by MACH'*s stripes*. In Fig. 15.1, a star cut from white paper has been pasted onto a dark cardboard disk. The disk is illuminated at a window or by a lamp and is rotated rapidly with a motor. The eye perceives three concentric circular zones: The innermost zone has the greatest brightness per surface area, and the outermost zone is the least bright. The middle zone provides a continuous transition between the other two. This is shown as a drawing in the lower part of Fig. 15.1.

However, we *see* – both on the rotating disk and in its photographic image, Fig. 15.2 – quite a different light distribution from what is in fact present. We *see* the inner, bright circle surrounded by a still brighter border. We *see* the dark ring bounded on its inner side by a still darker border. According to the strong impression given by our eyes, the most light would seem to come from the bright border, and

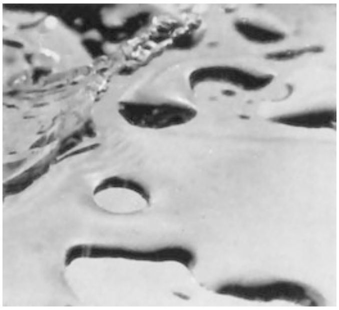


Figure 15.3 On "inverted seeing" (image of a surface which is partially wetted). Look at the image alternately as printed and then rotated by 180°.

the least light from the dark border. Every objective observer would be led inevitably to the wrong conclusion that the ambient light is most strongly reflected by the bright ring, and least reflected by the dark ring.

The *light* distribution sketched in the lower part of Fig. 15.1 can be observed in many arrangements and experiments. For this reason, these "MACH's stripes" have caused many an error in diverse physical observations.

Nevertheless, we should not hastily dismiss them as an "optical illusion". The phenomenon of MACH's stripes is of great importance for our visual sensory perceptions.

Think for example of reading dark printed letters on white paper. The lenses of our eyes by no means produce a perfect image. The outlines of the printed letters are not sharply imaged on the sensitive layer in our eyes, the retina. The transition from the dark letters to the bright background of the paper is washed out, as in a poorly-focussed photo. But our visual perceptions can compensate for this error with the help of MACH's stripes. The eye sees, figuratively speaking, a bright border at the boundary of the bright paper, and a dark border at the boundary of the dark letters in its image of the printed page. In spite of the fuzzy image on the retina, this gives us the impression of sharp outlines.

Another useful hint shows how our vision is decisively influenced by processes within our brains: Such "central processes" depend in a very complex way on the processes within the retina of the eyes. An example is "inverted seeing", as explained in Fig. 15.3 (an exchange of deep and high levels in an image). Keep in mind also what is said in Sect. 18.15 about the "depth of focus" in images. So much for these important phenomena which are generally typical of the operation of our visual sensory perceptive apparatus.

15.3 Physical Radiation Detectors. Direct Measurements of Radiant Power

Our eyes are by no means the only indicators for the radiation which is emitted by glowing objects: All bodies which are struck by the radiation will be *warmed*, that is, they receive energy. In the sun's radiation or the radiation from an arc lamp, we can feel this warming effect directly with the temperature sensors in our skin. The palm of the hand is especially sensitive.

This heating effect of the radiation offers a method for measuring its radiant power, that is the quotient of energy/time transmitted by the radiation. The principle is explained in Fig. 15.4. There, a metal plate is being irradiated by an incandescent lamp. The plate has been blackened with soot so that it absorbs practically all the incident radiation. A thermometer and an electric heater are also built into the plate.

We wait until a constant temperature has been established. Then equilibrium has been reached: During each time interval, just as much energy is brought to the plate by the radiation as it loses through heat conduction etc. Then we block off the radiation and adjust the heater current so that the same temperature is maintained; this requires a certain electrical power, that is a certain product of current and voltage, measured in volt · ampere = watt. This electrical power is equal to the previously-absorbed radiant power: Thus, we have calibrated the *radiometer*.C15.1

C15.1. For a quantitative description of the optical radiation in the visible wavelength range of light from around 400 to 750 nm (1 nm = 1 nanometer = 10^{-9} m), several additional, specialized quantities have been defined which take into account the perception of brightness by the human eye. For example, analogous to the *radiant intensity* with the unit watt/steradian, we have the *luminous intensity*, with the unit candela (one of the seven SI base units). These quantities will be discussed in detail in Chap. 29.

By comparing this calibrated but not very sensitive instrument with a more sensitive radiometer, e.g. a radiation thermocouple or thermopile, we can then calibrate the latter (Fig. 15.5).

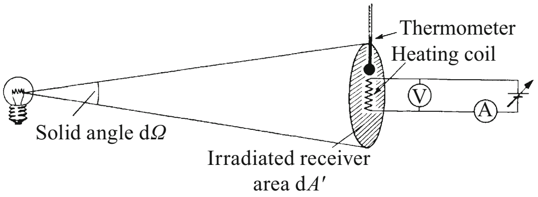


Figure 15.4 Calibration of an optical radiometer. The voltage of the current source can be adjusted. The radiant power $d\dot{W}$ is radiated into the solid angle $d\Omega$ and is then absorbed in the receiver area dA'; thus we define the *radiant intensity* of the lamp as emitter or source as $I = d\dot{W}/d\Omega$, and for the receiver area dA', the *irradiance* (or irradiation intensity) as $E_e = d\dot{W}/dA'$ (a detailed treatment of these quantities will follow in Chap. 19).

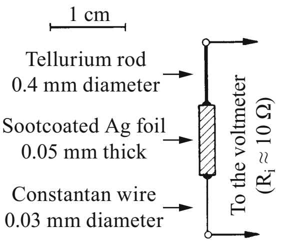


Figure 15.5 Schematic of a thermocouple (e.g. cubic SnTe-Constantan), into which a soot-covered Ag foil has been inserted for measuring the radiant power[C15.2]

C15.2. The thermocouple in Fig. 15.5 is a simple radiometer. The active (e.g. cubic SnTe-Constantan) junction is underneath the Ag foil, which serves as receiver for the radiation. The open circles indicate two other junctions where the Cu leads to the voltmeter are attached; they must be kept at the same constant temperature ("reference temperature") to avoid unwanted additional thermovoltages. A general summary of the use and design of thermocouples can be seen for example at www.msm.cam.ac.uk/utc/thermocouple/pages/ThermocouplesOperatingPrinciples.html. Today, radiometers often use a *thermopile*, a series of thermocouples connected together to increase sensitivity. An example of a commercial instrument is shown at www.kippzonen.com/Product/36/CA2-Laboratory-Thermopile#.V6xVelfuKrW.

15.4 Indirect Measurements of Radiant Power

In the case of radiometers which are based on the production of heat by the radiation (thermal radiometers), the incident radiant power is distributed over all the components of the absorbing object. The observed increase in its temperature corresponds only to the average energy increase of all the molecules in the detector. This limits the sensitivity of such radiometers. Much more sensitive radiometers are based on allowing all the radiant energy to be absorbed by only a small component, namely only by some fraction of the electrons which are a component of the radiometer. The electrons which receive the energy can be conveniently measured in the form of an electric current. This is the case for example with vacuum photocells (Fig. 15.6, left) (The photoeffect or photoelectric effect was described in the 13th edition of POHL's "*Optik und Atomphysik*", Chap. 14,), with photodiodes (Fig. 15.6, right) (cf. 21st edition of POHL's "*Elektrizitätslehre*", Chap. 27), with ionization chambers (Fig. 15.7), and with GEIGER-MÜLLER counters in their different forms (*ibid*., Chap. 20). In all of these devices, the measured electric currents are *proportional* to the absorbed radiant power. They thus provide an indirect measurement of the radiant power. Unfortunately,

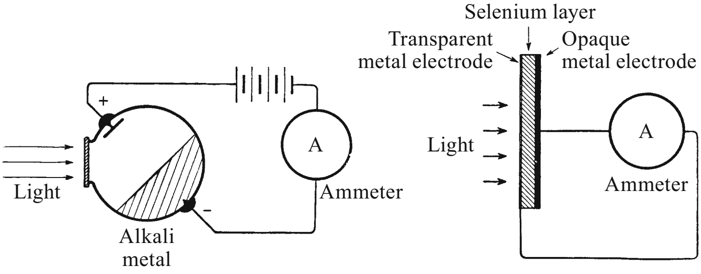


Figure 15.6 A vacuum photocell (left) and a photodiode (right). Both are convenient to use as radiometers in demonstration experiments, but they are unfortunately very selective. That means that their indicated values are indeed proportional to the radiant power, but they must be separately calibrated for every type of radiation.

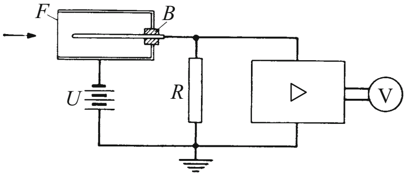


Figure 15.7 A gas-filled ionization chamber for detecting X-rays, in the form of a cylindrical condenser, used in connection with a DC amplifier and a voltmeter V ($U \approx 10^3$ V, $R \approx 10^9$ Ω). F is an aluminum foil entrance window for the radiation, and B is an insulator.

the proportionality constants depend on the type of radiation being measured. Their application therefore requires much more physical knowledge than in the case of a thermoelement. Where radiation measurement instruments (radiometers) appear in the illustrations of this book, we can generally take them to be thermocouples. When a more sensitive measurement is required, the necessary information will be provided in the description of the experimental setup.

Technical details have no place in this book. Nevertheless, we mention two points:
1. A very high sensitivity and broad applicability are exhibited by *photomultiplier* tubes, which are technically advanced vacuum photocells with a built-in amplifier (electron multiplier): The primary electrons which are ejected from the photocathode by light are accelerated by a first-stage voltage onto a metal plate (e.g. AgMg). They cause secondary electrons to be emitted there, whose number is a large multiple of the number of primary electrons. These secondary electrons are accelerated in turn by a second voltage stage onto another plate, ejecting tertiary electrons, and so forth through a number of stages (called ‘dynodes’), providing a very high overall amplification factor.
2. In order to make use of the convenient aspects of AC amplification, one uses a “chopper” to irradiate the instrument with light pulses (*intermittent light*). This has the added advantage that measurements can be carried out without requiring a darkened room: The constant current resulting from the background illumination is filtered out by the AC amplifier.

The Simplest Optical Observations

16

16.1 Light Beams and Light Rays

Physics is and remains an empirical science. In optics, as in other areas, observations and experiments have to provide the starting points for concepts and physical laws. It is reasonable to begin our treatment of optics, as in other fields, with the simplest experiences from daily life.

Every human being knows the difference between clear and hazy air, and between a clear and a murky liquid. Hazy air contains a large number of microscopic suspended particles, usually referred to as smoke, smog or dust. In a similar manner, liquids are clouded by small suspended particles. We can for example make clear water murky by adding a small amount of India ink, which contains very finely divided carbon powder, or with a few drops of milk, a suspension of microscopic particles of fat and casein **(Video 16.1)**.

Video 16.1:
"Polarized light"
http://tiny.cc/5dggoy
In this video, a plastic dispersion (styrofan) is used to make the light beam visible in a water-filled cuvette (see Fig. 24.4).

Room air is always more or less hazy; it usually contains many suspended dust particles. If necessary, tobacco smoke can be used to increase its haziness. Now, we carry out our first experiment using room air (Fig. 16.1): We employ a carbon-arc lamp in the usual sheet-metal housing as light source. The front end of the housing has a circular opening B which serves as the exit aperture for the light. Looking from the side, we can see a white, shimmering cone of light which extends from this opening far into the room. The light thus propagates within a cone (which is bounded by straight, dashed lines in the figure). It is called the *light beam*. This light beam has a large *opening angle* ω; it is determined by the opening B which serves as an aperture diaphragm. Travelling in beams with straight-line borders was listed in Vol. 1 as one of the basic properties of wave propagation (Sect. 12.6), provided that the wavelength is small compared to the diameter of the aperture (Fig. 16.2).

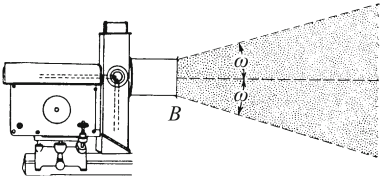


Figure 16.1 The visible trace of a light beam in hazy air (the dashed rays were drawn in later)

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_16

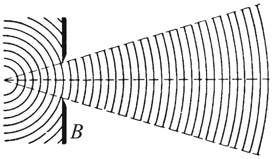


Figure 16.2 The propagation of mechanical waves as a beam with straight-line boundaries. The sketch shows water waves before and after passing through a wide opening (schematic, after Fig. 12.12 in Vol. 1).C16.1

C16.1. This is only a schematic sketch. In the case of mechanical waves, e.g. water waves, the boundaries of the beam can be only roughly discerned due to diffraction (see Vol. 1, **Video 12.2: "Experiments with water waves"** http://tiny.cc/tfgvjy).

The experiment shown in Fig. 16.1 illustrates the *visible trace* of a light beam in a hazy medium. The dust particles which are struck or illuminated by the light *scatter* a small fraction of it in all directions, and some portion of this scattered light reaches our eyes. Isotropic scattering from small particles is familiar from the mechanics of waves. We recall a stick standing up in the smooth surface of a pond: When water waves strike it, the stick becomes the source of circular "secondary" wave trains with propagate in all directions over the water surface (cf. Vol. 1, Fig. 12.17).

The further we move the opening away from the light source (the carbon arc) in Fig. 16.1 (arc discharges were discussed in the 21st edition of POHL's "*Elektrizitätslehre*", Chap. 18), the more narrow the light beam and the smaller its opening angle ω become. In the limiting case, the boundaries of the beam appear practically parallel as viewed from the side. We then speak of a *parallel-bounded beam of light*, or, for short, a *collimated beam*. In drawings, we represent a light beam in one of two different ways:

1. By two *boundary* rays (e.g. chalk marks or pencil strokes) at the sides of the beam. They define twice the opening angle, 2ω.

2. By a single ray representing the *beam axis*. It defines the direction of the light beam relative to some fixed direction.

We thus use the same techniques to represent light beams as we did for the cones or beams of mechanical waves (cf. Fig. 16.2). There, the rays in the drawing clearly represent lines normal to the wavefronts.

Only light *beams* can be *observed*. *Light rays* exist only on the blackboard or on paper. They are merely an aid to graphical and mathematical representation.

Later, we will demonstrate *curved* light beams experimentally in a corresponding mannerC16.2 and will represent them using curved lines or rays.

C16.2. See Sect. 27.11 and **Video 27.1: "Curved light beams"** http://tiny.cc/wfggoy.

For demonstrations to a large audience, we need very hazy air in order to make the trace of the light beam sufficiently bright. But we can avoid this problem; instead of hazy air, we use a cloudy liquid in a trough, or even more conveniently, a flat table with a matte finish. We can obtain the latter by painting a flat board with matte white or covering it with a sheet of white paper.

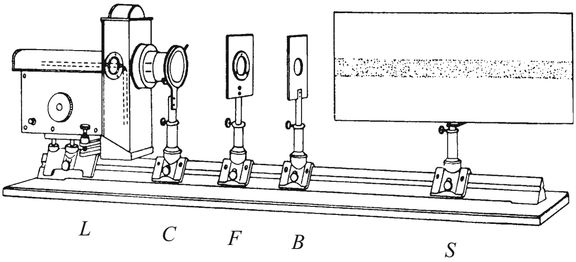


Figure 16.3 The visible trace of a collimated light beam along a white painted board S (B is a round aperture, F a red filter). To avoid having too great a distance from the light source with the associated problems, a lens C (called a *condenser* lens; cf. Sect. 18.3) with a focal length of around 7 cm (Sect. 16.7) is placed in front of the exit aperture of the light source.

> The "dust" in commercial white paints consists of a very fine powder made from a transparent material. Clear rock salt when powdered appears white. Clear ice gives white snow in the form of small particles. If "light" or "dark" beer is spread thinly in the form of fine bubbles, it appears as a white foam "head". White paper has a similar structure to white pigments; instead of a suspension of very finely powdered crystals, it consists of fine fibers which are tangled together and held by a layer of glue (cf. Sect. 26.10).

We can thus allow the light to shine at a glancing angle along a white painted board. Then we see its trace with a nearly blinding brightness. For the demonstration of collimated light beams, we make use of the convenient trick shown in Fig. 16.3. With this arrangement, we can also readily demonstrate a "colored"[1] light beam, for example a beam of red light; we need only place a red filter in front of the light source, for example a darkroom filter. *We will continue to work in the following with red-filtered light.*

For the light which we commonly encounter in daily life, that is the light of the sun, light from the sky, light from electric incandescent lamps, candles, gas lamps,[C16.3] or carbon-arc lights, we use the compact collective term *natural light*. The usual word "white" light is too vaguely defined.

16.2 Light Sources of Small Diameters

To demonstrate many optical phenomena, we require light sources with a small diameter and a strong luminous exitance (source brightness). The choice is limited.[C16.4]

Among them are the carbon tips (the "crater") of small arc lamps (diameter ≈ 3 mm), or the small arcs in high-pressure mercury lamps

C16.3. See Comment C1.1. in Vol. 1.

C16.4. Light sources which fulfill these conditions in an excellent manner are available today in the form of *lasers*. (The word "laser" is an acronym and stands for "Light Amplification by Stimulated Emission of Radiation". Described in the 13th edition of POHL's "*Optik und Atomphysik*", Chap. 14; see H.J. Eichler/J. Eichler, "Lasers" (Springer Verlag, Berlin 2003)). Some of the experiments described in this book can therefore also be demonstrated using a laser as light source (see e.g. **Video 27.1: "Curved light beams"**). In many cases, the use of a laser however provides no particular advantage, so that an aperture or slit illuminated by an arc lamp and condenser lens has by no means lost its usefulness as a light source.

[1] "Colored light" or "red light" is in terms of language at the same level as a "high note". Both expressions are justifiable only because of their convenient brevity.

(diameter $\approx 0.3\,\text{mm}$)[2]. In general, the boundaries of the source in such lamps are not sufficiently sharp. Therefore, instead of a lamp as light source, we often employ a circular aperture illuminated from behind, or a slit with straight edges. For the illumination from behind, we place a lens of short focal length between the lamp and the aperture: a *condenser* lens. One of many examples can be seen in Fig. 16.3. The details of a proper illumination will be given later in Fig. 18.12.

16.3 The Fundamental Facts of Reflection and Refraction

Making use of the experimental aids described above, we now begin by recalling school physics and the laws treated in some detail in Vol. 1 (Sect. 12.7): the law of reflection and the law of refraction. We employ the setup illustrated in Fig. 16.4. A thin red light beam I falls at a slant from the upper left through the air onto the planar, polished surface of a glass block B. At the surface, it is *split* into *two partial beams* II and III. One of them, beam II, is reflected towards the upper right. After the reflection, the rays as drawn seem to originate from the "virtual" intersection point L', the "mirror image" of the object point. The other beam, III, enters the glass block, changing its direction of propagation at the surface; it is *refracted*. All the rays drawn in the figure lie in the same plane, the "plane of incidence" (the plane of the page). Three of these rays belong together in each case; they make the three adjacent angles α, β and α' with the "axis of normal incidence" N. In Fig. 16.4, these angles are shown for the center axes of the beams; the boundary rays are left off for clarity. The set of angles obeys the *law of reflection*:

$$\alpha = \alpha' \,, \tag{16.1}$$

and for the transition of the light from the *air* into the material B (glass), the *law of refraction* (SNELL's law)C16.5

$$\frac{\sin\alpha}{\sin\beta} = \text{const} = n_\text{B} \,. \tag{16.2}$$

n_B, often written without the subscript, is called the *index of refraction* of the material B. Some numerical values are collected in Table 16.1. In comparing two materials, the one with the higher index of refraction is referred to as the more "optically dense" material.

In Fig. 16.4, we show a boundary surface between air and glass. Instead of this, we could employ a boundary surface between two

C16.5. WILLEBRORD SNELL (1580–1626), a Dutch mathematician. In Sect. 12.8, the index of refraction n for electromagnetic waves was introduced; it thus holds for light, as $n = c_\text{vacuum}/c_\text{matter}$ (A detailed description for light follows later in Chap. 25).

[2] Even this diameter is still very large compared to the wavelength of visible light (Sect. 16.9). In acoustics, in contrast, we can easily make the apertures of radiation sources (e.g. of pipes and whistles) smaller than the wavelength of the sound waves.

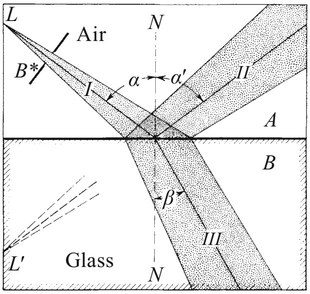


Figure 16.4 Demonstration of reflection and refraction of a light beam at the planar surface of a glass block (flint glass). The block is standing in front of a matte white screen, and its back face is also ground to a matte finish. (Red filter light, L; the light source has a small diameter, and B^* is its aperture diaphragm)

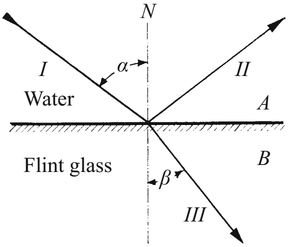


Figure 16.5 Reflection and refraction at the planar boundary surface between two materials, water (A) and flint glass (B), with different indices of refraction, n_A and n_B (red filter light; only the center axes of the light beams are drawn)

arbitrary transparent materials A and B (with the indices of refraction n_A and n_B), for example, as in Fig. 16.5, between water and flint glass. The law of reflection is unchanged, while in the case of refraction, we find for the transition from material A into material B:

$$\frac{\sin\alpha}{\sin\beta} = n_{A\to B} = \frac{n_B}{n_A}\,, \tag{16.3}$$

e.g. $n_{\text{water}\to\text{flint glass}} = \frac{1.60}{1.33} = 1.20$; cf. Table 16.1.

A comparison of Eqns. (16.2) and (16.3) yields $n_A = n_{\text{air}} = 1$. We have thus defined the index of refraction of a material (as in general and expedient usage) in terms of the transition of the light from room air into that material. For the transition from vacuum → material, the

Table 16.1 The indices of refraction of some materials

For the transition of red filter light ($\lambda \approx 650$ nm at 20 °C), from air into	Index of refraction $n =$
Fluorspar	1.43
Quartz glass	1.46
Light crown glass (lead-free silicate glass)	1.51
Rock salt crystal	1.54
Light flint glass (silicate glass with ≈ 25 wt.-% PbO)	1.60
Heavy flint glass (silicate glass with ≈ 40 wt.-% PbO)	1.74
Diamond	2.40 (!)
Water	1.33
Carbon disulfide	1.62
Methyl iodide (iodomethane)	1.74

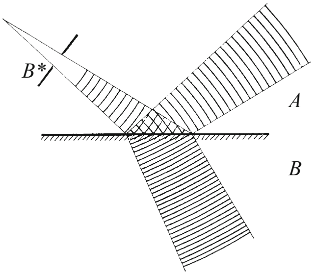


Figure 16.6 Reflection and refraction of mechanical waves (e.g. water waves) at the boundary between two materials with different wave velocities (above higher than below; thus shorter wavelength below) (schematic)

corresponding indices of refraction would be about 0.3 thousandths higher. Thus, taking this transition as the basis of the definition, room air would have the index of refraction $n_{\text{vacuum}\to\text{air}} = 1.0003$.

For *mechanical* waves, we observed reflection and refraction in the form sketched in Fig. 16.6. The rays drawn continue as normals to the wavefronts after reflection. Quantitatively, we find:

$$\frac{\lambda_A}{\lambda_B} = \frac{n_B}{n_A} \quad \text{or} \quad \lambda_B = \frac{\lambda_A}{n_{A\to B}}\,. \tag{16.4}$$

We will have occasion to apply this equation later to light, as well.

Figure 16.7 describes the same experiment as Fig. 16.5, however here for the special case of a collimated light beam. In addition to the two light rays at the edges of the beam, two cross-sections are drawn in as the intersection lines 1 and 2. In a wave picture, these are *wavefronts*, e.g. they represent crests of the waves.

From this sketch, we can read off the following results:

$$\frac{s_A}{s_B} = \frac{\sin\alpha}{\sin\beta} = \frac{n_B}{n_A}$$

or

$$s_A \cdot n_A = s_B \cdot n_B\,. \tag{16.5}$$

In words: Between two cross-sections of a light beam, the product of path and index of refraction, called the *optical path length*, is constant. This is FERMAT's principle.C16.6

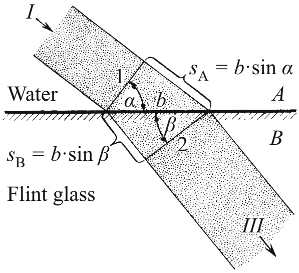


Figure 16.7 The definition of the optical path of a collimated light beam. The reflected light beam is not shown, to keep the drawing simple and clear.

C16.6. In equal time intervals Δt, a light beam passes through a distance $s_A = c_A\Delta t$ or $s_B = c_B\Delta t$. It follows from this, according to Comment C16.5, that $\Delta t = (1/c_{\text{vac}})n_A s_A$ or $\Delta t = (1/c_{\text{vac}})n_B s_B$. The quantity $n \cdot s$ is thus a measure of the time required by the light in order to pass through the distance s with the velocity c_{matter} (where $1/c_{\text{vac}}$ is a constant of proportionality). This quantity, the *optical path length*, will repeatedly play a role in the following sections. We mention an important application here: For the path which is chosen by a light wave in order to go from a point P to another point P', it is found that the optical path length, in general the path integral

$$\int_P^{P'} n\,\mathrm{d}s\,,$$

is a *minimum*, so that the light covers the distance most rapidly along this path (this holds also in inhomogeneous media). This is FERMAT's principle (PIERRE DE FERMAT, 1601–1665, French mathematician). See for example Max Born and Emil Wolf, *Principles of Optics* (Pergamon Press, 4th edition (1970)), Sect. 15. (available online – cf. Comment C25.11).

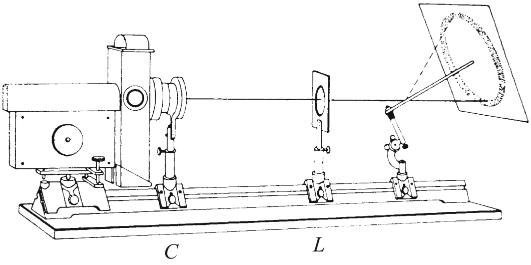


Figure 16.8 The reflection cone formed by reflection of light from the surface of a cylindrical glass rod (C is a condenser lens; its mount carries an iris diaphragm about 8 mm in diameter. L is a lens of focal length $f = 20\,\text{cm}$) **(Video 16.2)**

To illustrate the *law of reflection* (16.1), we mention a practically-important but little-known special case: In Fig. 16.8, a thin light beam strikes the smooth surface of a cylindrical rod at a grazing angle. The light reflected from the surface forms a cone. The axis of the cone coincides with the axis of the rod; thus, a screen perpendicular to the rod shows a circular line at the intersection with the cone of the reflected light. The direction of the incident light beam lies within the surface of the cone. The more steeply the incident beam strikes the rod, the larger is the opening angle of the cone.

Video 16.2:
"Reflection cones"
http://tiny.cc/qdggoy
In the video, a reflecting stainless steel tube is held by hand in a light beam. Depending on the angle between the tube's axis and the light beam, reflection cones are formed, whose circular sections can be seen on the wall of the lecture room. They change depending on the position of the tube. At an angle of 90°, the section becomes a linear band of light.

Knowledge of this type of reflection is important for example in the investigation of rod-shaped formations using dark-field illumination, e.g. with an optical microscope (Sect. 18.12) or an electron microscope. It is also important for diffraction of X-rays by crystal lattices (Sect. 21.14) and for the explanation of atmospheric *halo* phenomena, in which a ring touches the image of a star or planet on the outside (for references, see Comment C21.1). A beautiful collection of haloes can also be seen at https://en.m.wikipedia.org/wiki/Halo_(optical_phenomenon) .

16.4 The Law of Reflection as a Limiting Case. Scattered Light

According to the illustration shown in Fig. 16.4, the reflected light should be restricted to beam *II*, that is to a three-dimensional cone with its apex at L'. This description however holds only for an ideal limiting case: In reality, we can see the point at which the light beam I intersects the surface from any arbitrary direction. Thus, some portion of the incident light must be "scattered" in a diffuse manner in all directions and thus reaches our eyes. This *scattered light* is a cause of annoyance to physicists and engineers, as a disturbing source of errors; but it is a blessing for fathers: Without the scattered light, their children would continually run into plate glass doors and windows.

"… all objects which themselves do not emit light become visible to us only through scattered light".

For all objects which themselves do not emit light become visible to us only through scattered light.[C16.7]

C16.7. this important statement should be particularly emphasized here. The decisive role of scattered light for "seeing" objects is often not noticed by the novice.

Scattered light is produced in the main by imperfections in the smooth surface, for example due to dust particles, imperfect polishing and inhomogeneities in the material. The diameter of dust particles is seldom less than around 10 μm. Light scattering thus occurs primarily through *reflection* by innumerable small, randomly-oriented mirror surfaces. Therefore, this type of light scattering is expediently referred to as *diffuse reflection*. The scattered light vanishes almost completely when the reflecting surface is nearly perfect, produced without mechanical treatment; an example is the freshly-prepared surface of pure mercury, or the recently-cleaved surfaces of mica crystals.

The dust particles which fall onto a mercury surface can be burned off by waving the flame of a BUNSEN burner over it.[C16.8]
Mica sheets must be cleaved both on their upper and their lower sides.

C16.8. Mercury surfaces are used in particular to produce concave parabolic mirrors. The shape is obtained through rotation (Vol. 1, Fig. 9.3). See e.g. the note by R.F. Wuerker in *Physics Today*, July 2004, p. 82.

16.5 Total Reflection

Total reflection is also treated in detail in Vol. 1.[C16.9] For light, we demonstrate it with the arrangement sketched in Figs. 16.9 and 16.10. The light beam passes from the more optically dense material (*B*) to the less dense medium (*A*), this time, exceptionally, from right to left. The corresponding angles are again drawn in only for the central axes of the light beams. We can reach two conclusions based on these figures:

C16.9. See Vol. 1, Sect. 12.9 and **Video 12.2, "Experiments with water waves"** http://tiny.cc/tfgvjy (at 5:30 minutes).

1. The refracted light beam *III* propagates at a larger angle to the interface normal *N* than the incident beam *I*. Experimentally, we find

$$\frac{\sin\alpha}{\sin\beta} = n_{\mathrm{B}\to\mathrm{A}} = \frac{n_\mathrm{A}}{n_\mathrm{B}} = \frac{1}{n_{\mathrm{A}\to\mathrm{B}}}\,. \tag{16.6}$$

The axes of the incident and the refracted light beams show the same patterns in Figs. 16.4 and 16.9; only the optical path is reversed in the two figures.

2. At large angles of incidence α, there is no longer a refracted beam *III*. All of the incident light is reflected: *Total reflection* occurs in Fig. 16.10. Quantitatively, the angle β cannot become larger than 90°, that is, its sine cannot become larger than 1 in Eq. (16.6). Thus, the formula

$$\sin\alpha_\mathrm{T} = \frac{n_\mathrm{A}}{n_\mathrm{B}} = \frac{1}{n_{\mathrm{A}\to\mathrm{B}}} \tag{16.7}$$

determines the "critical angle" α_T for total reflection. The critical angle α_T corresponds in the optically less dense medium to a *grazing* beam, i.e. to a beam which propagates parallel to the interface (compare Vol. 1, Fig. 12.25).

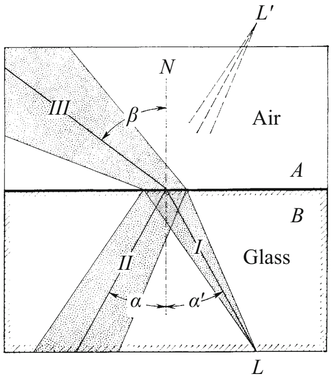


Figure 16.9 Reflection and refraction of a light beam in passing from an optically more dense to a less dense medium (red-filter light). The angle of incidence is again denoted by α.

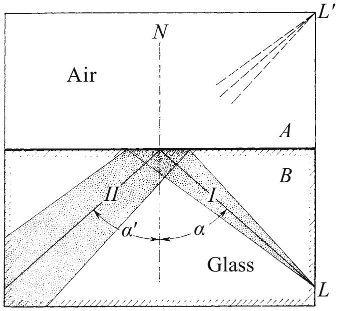


Figure 16.10 The continuation of Fig. 16.9. When the angle of incidence α is increased, the refracted light beam is no longer present: Total reflection occurs.

Total reflection is a favorite object for demonstration experiments; there are many setups for showing it. The best-known is a gimmick, using a jet of water as a light guide (a “luminous fountain”).C16.10 In nature, total reflection can frequently be observed with air bubbles under water; think of the bright, silvery shining bubbles on the sides of water beetles.

The critical angle for total reflection can be rather precisely determined by a variety of methods. This fact is used in the construction of *refractometers*.C16.11 These are apparatus for the rapid and convenient measurement of the index of refraction (usually of a liquid sample); they are popular with chemists and medical researchers. An example is described by Fig. 16.11.

Total reflection can occur at the interface between two media with a very small difference in their indices of refraction. The light beam must be incident at a grazing angle, i.e. α must be nearly 90°. In this way, we have seen how sound waves can be reflected from the interface between warm and cool air (Vol. 1, Fig. 12.56). The corresponding effect is also seen with light beams (Fig. 16.12): A collimated light beam enters at a grazing angle from below into a box which is electrically heated. The inside surfaces of the box are blackened. When the heater is turned on, the box fills with hot air; some of it expands over the rim of the box, while the rest forms a rather flat interface (diffusion boundary as a substitute for a surface; cf. Vol 1,

C16.10. Light guides made of glass fibers (“fiber optics”) work on the same principle; they are used extensively today for optical-digital data transport, for example in telecommunications, and for endoscopic examinations (technical details can be found e.g. in H. Kogelnik, “*Optical Communications*”, in the *Encyclopedia of Applied Physics*, Vol. 12, p. 119 (VCH Publishers, 1995).

C16.11. Refractometers are commercially available in many forms. Along with the measurement of indices of refraction, they can be used for example to indicate the alcohol or sugar content of solutions directly as a digital output.

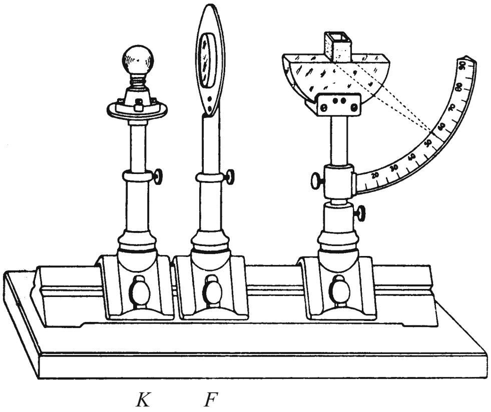


Figure 16.11 A refractometer which is suitable for demonstration experiments. A thick, semicircular glass plate with a known, large index of refraction n_B carries a glued-on rectangular glass cuvette into which the liquid of unknown index of refraction n_A is filled. At left, parallel to the flat face of the glass plate and about 30 cm away, is a lamp K with a red filter F in front of it. The light which passes through the liquid and enters the glass plate at a grazing angle appears on the protractor scale as a thin red beam whose right vertex as seen by the observer is quite sharply defined. The critical angle α_T can thus be rather precisely read off, and n_A is then computed from Eq. (16.7), or else the angular scale is calibrated directly in units of the index of refraction. The round glass plate acts as a cylindrical lens (Sect. 16.7). This is indicated in the figure by two dashed convergent rays.

Sect. 9.9). This interface between hot and cool air acts like a fairly planar mirror. Strong drafts perturb the experiment and should be avoided.

This total reflection by a warm layer of air is often observed in nature. The hot ground in a desert region or a hot asphalt road surface warms the air just above it. A traveller sees a mirror image of the bright sky at a grazing angle, sometimes containing the mirror image of some object near the horizon (*fata morgana* or *mirage*). The totally-reflecting boundary layer usually appears as if it were the surface of a pool of water.

In physics, total reflection at grazing incidence plays a significant role in spectral apparatus ("monochromators") for X-ray light (Sect. 22.6).

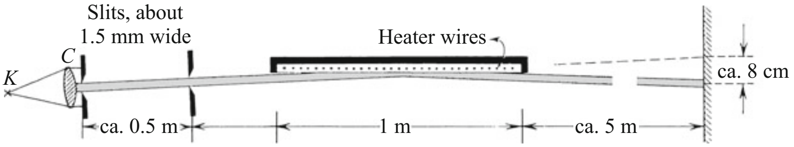


Figure 16.12 Total reflection of a collimated light beam at the interface between hot and cool air (the beam is about 2 cm thick at its right-hand end; K is the crater of an arc lamp (light source point))

16.6 Prisms

Prisms represent important applications of the law of refraction in optics and metrology. In Fig. 16.13, we see the two planar faces of a prism which make an angle γ with each other, the “angle of refraction”. Perpendicular to the two faces, the plane of the page forms the *principal plane*. Within the principal plane, a collimated beam of light passes through the prism; only its central ray is drawn in the figure. Refraction at the two faces of the prism changes the direction of the beam by the angle of deflection δ. Quantitatively, we can apply the equation

$$\sin\alpha = n\,\sin\beta \tag{16.2}$$

after some rearrangements (for the relation between δ, α, β, and γ, see **Exercise 16.1**)

$$\tan\left(\beta - \frac{\gamma}{2}\right) = \tan\frac{\gamma}{2}\cdot\frac{\tan\left(\alpha - \dfrac{\delta+\gamma}{2}\right)}{\tan\left(\dfrac{\delta+\gamma}{2}\right)}\,. \tag{16.8}$$

The *minimum deflection* can be determined experimentally when the collimated light beam passes *symmetrically* through the prism, as in Fig. 16.13, right. Then we find

$$\beta = \frac{1}{2}\gamma \quad \text{and} \quad \alpha = \frac{1}{2}(\delta+\gamma)$$

(here, α is the angle of incidence to the normal, β is the angle of the refracted beam, γ is the angle of refraction (apex angle) of the prism, and δ is the angle of deflection of the light beam).

Then we obtain from Eq. (16.2)

$$n = \frac{\sin\frac{1}{2}(\delta+\gamma)}{\sin(\gamma/2)} \tag{16.9}$$

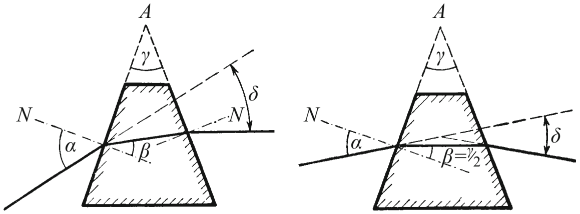


Figure 16.13 The deflection of a monochromatic beam (the ray drawn indicates the central axis of the light beam) by a prism with a non-symmetric optical path (*left*) and with a symmetric path (*right*). The dashed lines marked ‘*N*’ are surface normals to the prism faces. The line perpendicular to the principal plane (the plane of the paper) at the apex point *A* is called the *refracting edge* of the prism.

and

$$n = \frac{\sin\alpha}{\sin(\gamma/2)} \,. \tag{16.10}$$

These two equations can be used to determine the index of refraction n. We measure either δ or α.

In the limiting case of small angles of refraction γ, we can replace the sine and tangent functions in Eqns. (16.8) and (16.9) by their arguments (angles). Then we find for both the non-symmetric and for the symmetric optical paths a deflection angle of

$$\delta = (n-1)\gamma \,, \tag{16.11}$$

i.e. the deflection angle δ is proportional to the angle of refraction γ of the prism.

16.7 Lenses and Concave Mirrors. The Focal Length

A divergent beam of water waves which is defined by an opening S can be made convergent by means of a lens (Fig. 16.14). One thus obtains a narrow "waist" in the beam of waves within a short region, denoted succinctly as an "*image point*" L'. Analogously, in optics, we can allow a divergent light beam to fall on an aperture S and can convert it into a convergent beam (Fig. 16.15) by placing a lens in the aperture. In this way, we can form an "image" of a pointlike light source L (the "*object point*"). In Fig. 16.15, the beam axis and its two limiting boundary rays are drawn. The *aperture* which defines the beam is at the same time the mount S of the lens. The center point of the aperture thus lies here on the (dot-dashed) symmetry axis of the lens, the *optical axis*. In this case, the axis of the light beam has a special name, the *principal ray*.

The quantitative treatment of lenses begins with cylindrical lenses. If we wish to form an image of each *point* in the object with a cylindrical lens as an image *point*, then we must not employ three-dimensional light beams, but rather practically *two-dimensional* or planar light beams. This means that we must use an aperture which

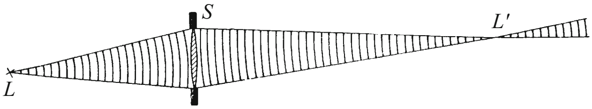


Figure 16.14 A lens converts a divergent beam of mechanical waves into a convergent beam (schematic, as in Fig. 12.20 in Vol. 1; see also **Video 12.2: "Experiments with water waves"** http://tiny.cc/tfgvjy)

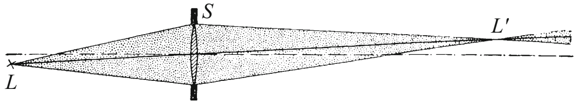


Figure 16.15 A lens converts a divergent beam of light which is defined by the aperture (lens mount) S into a convergent beam (L' is a real image point, schematic)$^{C16.12}$

C16.12. To distinguish it from *virtual* image points, which will be introduced later, we refer here to a *real* image point.

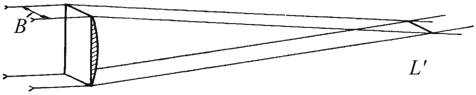


Figure 16.16 Imaging of a distant object point by a cylindrical lens as an image streak L'. One must limit the width B of the incident light beam using a slit in order to convert the "image streak" into an "image point".

takes the form of a narrow *slit* which is perpendicular to the cylinder axis of the lens. Continuing the experiment, we open the aperture until it reaches the width B marked in Fig. 16.16 with a double arrow. Then a cylindrical lens produces for each object point L an image *streak* L' rather than an image *point*. *Only with two crossed cylindrical lenses with the same radius of curvature can we obtain the effect of a spherical lens*; i.e. for each object point L they produce an image *point* L' (Fig. 16.17a) and thus yield good (sharply focussed, well resolved) images. Two crossed cylindrical lenses with *different* radii of curvature produce two perpendicular image *streaks* L' and L'' at different spacings, instead of an image *point* for each object point (Fig. 16.17b; this is called *astigmatism*, cf. Sect. 18.5).

Starting from the cylindrical lens, we can relate the action of a lens to the action of prisms. We restrict ourselves to a nearly flat cylindrical lens, i.e. with a very limited curvature ("*thin lens*", Fig. 16.18) and a *light beam which is very narrow in both its dimensions and close to*

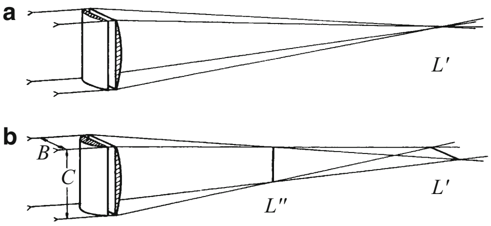


Figure 16.17 Imaging of a distant object point by two crossed cylindrical lenses (a) with the same radius of curvature: One obtains a single image point L'; (b) with different radii of curvature: One obtains two separate image streaks L' and L'' (cf. Sect. 18.5, Astigmatism). With a slit, one can reduce either the width B or the width C of the incident light beam. In the first case (width B small), we convert the image streak L', and in the second case (width C small), we convert the image streak L'' into an "image *point*".

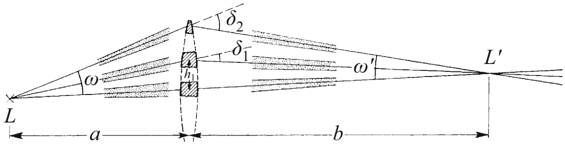


Figure 16.18 The relation between the action of a lens and the action of prisms. The radii of curvature of the two faces of the lens are denoted in Eq. (16.12) as r_1 and r_2.[C16.13]

C16.13. This is a good example of FERMAT's principle (Comment C16.6): Although the geometric path length depends on the angle of deflection δ, the *optical* path lengths are the same. Thus, the light can use all of the paths sketched in the figure to pass from L to L'.

the optical axis (called *paraxial rays*). (Unfortunately, in the drawings, we have to exaggerate the opening angles ω and ω' of the light beam for clarity!) This light beam is then divided up into small sub-beams as shown in Fig. 16.18, and only the central axis ray of each sub-beam is considered. At the same time, we divide the lens up into a series of prisms which are one above the other at altitudes h above the optical axis.

We thus arrive at the well-known *lens formulas* which are valid only in the limiting case of *thin light beams close to the optical axis* ("paraxial rays"):

$$(n-1)\left(\frac{1}{r_1}+\frac{1}{r_2}\right)=\frac{1}{f'}\,, \qquad (16.12)$$

$$\frac{1}{a}+\frac{1}{b}=\frac{1}{f'}\,. \qquad (16.13)$$

In these equations, f' is termed the image-side *focal length* (Fig. 16.19, left). Equation (16.12) is called the "lens-maker's equation", and Eq. (16.13) is the "imaging formula".

The derivation of Eqns. (16.12) and (16.13) is illustrated by Fig. 16.20. There, the lens is thick and has strongly curved faces, in order to make space for the large number of necessary symbols. For the small shaded triangle with the outer (supplementary) angle δ, we have

$$\delta=\varphi_1+\varphi_2=(\alpha_1-\beta_1)+(\alpha_2-\beta_2)\,. \qquad (16.14)$$

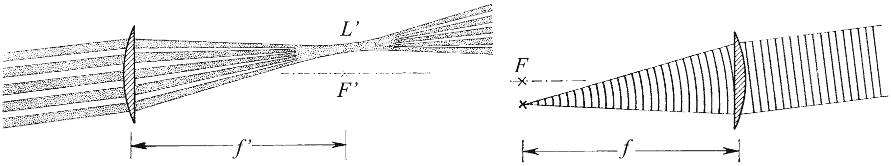


Figure 16.19 The definition of the image-side focal length f' (left) and the object-side focal length f (right). The former is demonstrated using a series of collimated light beams. They all originate at the same distant point L on the object. They are produced by subdividing a wide collimated light beam using a lattice aperture.

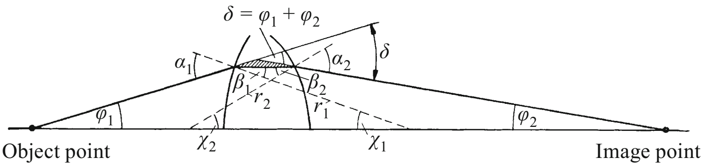


Figure 16.20 The derivation of Eqns. (16.12) and (16.13)

Then, according to the law of refraction,

$$\frac{\sin\alpha_1}{\sin\beta_1} = \frac{\sin\alpha_2}{\sin\beta_2} = n \tag{16.2}$$

or, for small angles, approximately

$$\alpha_1 = n\beta_1 \quad \text{and} \quad \alpha_2 = n\beta_2\,. \tag{16.15}$$

We then obtain from Eq. (16.14)

$$\delta = \varphi_1 + \varphi_2 = (n-1)(\beta_1 + \beta_2)\,. \tag{16.16}$$

Furthermore, the large triangle with angles χ_1 and χ_2 and the small triangle with angles β_1 and β_2 have the same apex angle; therefore, $\beta_1 + \beta_2 = \chi_1 + \chi_2$ and Eq. (16.16) takes on the form:

$$\delta = \varphi_1 + \varphi_2 = (n-1)(\chi_1 + \chi_2)\,. \tag{16.17}$$

Note that the right side of Eq (16.17) is simply an application of the prism formula, Eq. (16.11). Now, introducing the altitude h of the common apex in Fig. 16.20, we find $\tan\varphi_1 = h/a$ and $\tan\varphi_2 = h/b$, and also $\sin\chi_1 = h/r_1$ and $\sin\chi_2 = h/r_2$. With the small-angle approximation ($\tan\varphi \approx \varphi$, $\sin\chi \approx \chi$), we then rewrite Eq. (16.17):

$$\delta = \frac{h}{a} + \frac{h}{b} = (n-1)\left(\frac{h}{r_1} + \frac{h}{r_2}\right)\,; \tag{16.18}$$

or, dropping the leftmost equation and cancelling the common factor h,

$$(n-1)\left(\frac{1}{r_1} + \frac{1}{r_2}\right) = \frac{1}{a} + \frac{1}{b}\,. \tag{16.19}$$

Now we consider the case that the object is very distant, i.e. $a \to \infty$, $1/a \to 0$. Then the image will be formed in the focal plane at a distance $b = f'$ from the lens plane (cf. Fig. 16.19, left), so that $1/b = 1/f'$. Substituting these expressions into Eq. (16.19) leads immediately to Eq. (16.12), and comparing again to (16.19) gives (16.13).

The distances a and b and the focal length f' are measured *tentatively* from the central plane of the lens (more details will be given in Sect. 18.2). The set of all the image points of all the distant object points make up the image-side *focal plane*. Its intersection with the optical axis defines the image-side focal point F'.

In a corresponding manner, we define the object-side focal plane, the object-side focal point F, and the focal length f; cf. Fig. 16.19, right. The light rays which originate from the point L on the object-side focal plane (and are divergent on that side of the lens) form a collimated

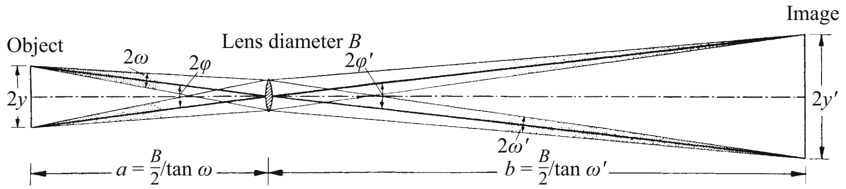


Figure 16.21 The imaging of an extended object by light beams originating from its individual object points. ω and ω' are called the object-side and the image-side opening angles. φ and φ' are the inclination angles of the principal rays, which are equal here (see also Comment C18.1).

beam after passing through the lens. For the comparison to mechanical waves, several wave crests are drawn in as transverse lines. For lenses in air (or in general with the same material on both sides), the object-side and the image-side focal lengths are the same.

Practitioners call the reciprocal of the focal length the *optical power* of the lens, i.e. power $P = 1/f$. The unit is $1\,\mathrm{m}^{-1} = 1$ diopter (analogously to $1\,\mathrm{s}^{-1} = 1\,\mathrm{Hz}$). A lens of power $1/f = 3$ diopter $= 3\,\mathrm{m}^{-1}$ thus has a focal length of $f = 0.33$ m. When several lenses are placed one behind the other, their powers add (approximately).C16.14

C16.14. The exact expression for the power of two (thin) lenses at a spacing d is given by the ‘Gullstrand formula’; see http://hyperphysics.phy-astr.gsu.edu/hbase/geoopt/Gullstrand.html. It can also be used to calculate the power of *thick* lenses (http://hyperphysics.phy-astr.gsu.edu/hbase/geoopt/gullcal.html).

Formation of the image of an extended object can be traced back to imaging of each of its individual points by a single light beam. This is illustrated in Fig. 16.21 for the uppermost and lowest points of an object. For many purposes, it suffices to sketch the *principal rays* which are shown here as heavy lines[3] (e.g. as in Fig. 18.25). From Fig. 16.21, we can read off the frequently-used relationsC16.15

C16.15. The image scale or magnification as defined here should not be confused with the increase in the viewing angle introduced in Sects. 18.10, 18.11 and 18.13 for optical instruments.

$$\text{Image scale} = \frac{\text{Image size } 2y'}{\text{Object size } 2y} = \frac{\text{Image distance } b}{\text{Object distance } a} \approx \frac{\tan\omega}{\tan\omega'} \tag{16.20}$$

(ω is the object-side opening angle, ω' the image-side opening angle). For angular units, see the footnote at the end of Chap. 17.

Furthermore, we have:

$$\text{Image size } 2y' = \text{Image distance } b \cdot 2\,\tan\varphi\,, \tag{16.21}$$

or, for small angles,

$$2y' = b \cdot \tan 2\varphi \tag{16.22}$$

(φ is the angle of incidence between the principal ray and the optical axis).

[3] We repeat: The *principal ray* is the name of the axis of the light beam in the case that the midpoint of the beam (in Fig. 16.21, this is the lens mount) lies on the symmetry axis of the lens (Fig. 16.15).

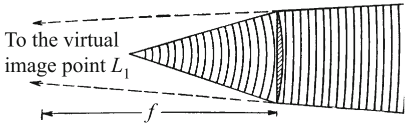


Figure 16.22 The object point lies inside the object-side focal plane. The lens only reduces the divergence of the beam.

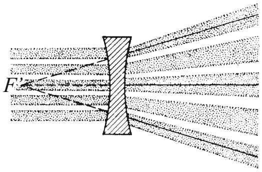


Figure 16.23 The action of a concave lens (F' is the virtual image-side focal point)

An example of Eq. (16.22): The disk of the sun has an angular diameter of $2\varphi = 32'$ (arc minute), or 0.53°. Its image is at a distance of $b = f$ behind the lens, that is $2y' = \tan 32' \cdot f = 9.3 \cdot 10^{-3} \cdot f$. A lens with a focal length of 1 m produces an image of the sun with a diameter of $2y' = 9.3$ mm.

A light beam from an object point L *inside* the object-side focal plane (Fig. 16.22) is not made convergent by the lens (i.e. it is not 'focussed'), but rather its divergence is only reduced. The dashed backwards extrapolations of the two boundary rays drawn lead to the *virtual* image point L_1. For the comparison with mechanical waves, a number of wave crests are also indicated in Fig. 16.22.

Concave lenses exhibit no essentially new properties.[C16.16] They increase the divergence of the light beam. Figure 16.23 illustrates this for the case of a collimated beam of light which is incident from the left. It also provides the definition of the image-side (virtual) focal point F'. Equations (16.12) and (16.13) remain valid with a proper choice of the signs.

C16.16. Concave lenses, or more generally, lenses which are thinner at their centers than at their outer rims, are also termed *diverging lenses*. In contrast, convex lenses, or more generally, lenses which are thicker at their centers than at their outer rims, are called *converging lenses*.

Concave mirrors are mainly important for physics and astronomical applications in a particular mode of employment: The object or image points are in the focal plane, near the symmetry axis (optical axis) of the mirror, and the opening angle of the light beam is only moderately large. The action of the concave mirror can then be found from simple geometrical considerations employing the law of reflection. The focal length of the concave mirror is equal to half its radius of curvature R (Fig. 16.24).

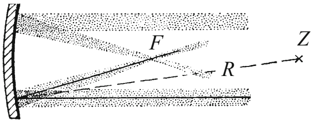


Figure 16.24 The action of a concave mirror (**Exercise 16.2**)

16.8 The Separation of Parallel Light Rays by Image Formation

Many optical phenomena take their simplest form when they are observed using collimated (i.e. *parallel*) light beams. In experiments of this type, one is often dealing with a splitting of a single parallel light beam into two or more beams. In the simplest case, we observe the scheme shown in Fig. 16.25. A collimated light beam is incident from the left and passes through some sort of apparatus *G*. There, it is separated into two parallel light beams. But the separation is not complete, the beams still strongly overlap.

How can we obtain a sufficiently effective separation of two beams? From the geometrical appearance, we would be inclined to say: First, we must make the cross-sectional areas of the collimated beams small; and second, we shift the plane of observation in Fig. 16.25 further to the right.

Both of these suggestions presuppose a strictly parallel form of the beam. The beams must not become unfocussed, neither on limiting their cross-sections nor at a large distance from *G*, nor should they be broadened. However, it is impossible to fulfill these conditions with real light beams. All so-called collimated or parallel beams are in fact somewhat divergent. Among the various reasons for this, we mention here only one, namely the finite diameter of all available light sources.C16.17

C16.17. For a discussion of the feasibility of producing truly parallel, collimated light beams, see also Sect. 19.6.

The incomplete separation of the beams can be eliminated by using a lens (Fig. 16.26). It converts every collimated light beam into a convergent beam. The observation is carried out in the image plane, where the "waist" of the beams is narrowest.

For demonstration experiments, an approximation is always sufficient. As in Fig. 16.27, we set a lens *in front of* the apparatus *G*. The light is incident on the lens as a divergent beam. The image plane is shifted far to the right, usually several meters. Then the light beams which converge to the image points are very slender, and the apparatus *G* receives nearly collimated light beams.

Figure 16.25 Incomplete separation of two parallel light beams after passing through some sort of apparatus *G*

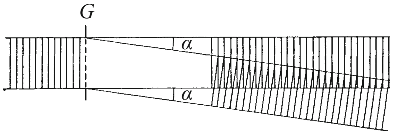
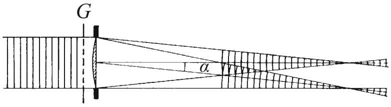


Figure 16.26 The unwanted overlap is eliminated after concentrating the two beams into one image point each



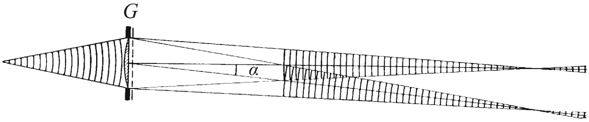


Figure 16.27 A simplification of the arrangement sketched in Fig. 16.26, which is adequate for demonstration experiments. For comparison with beams of mechanical waves, in Figs. 16.25–16.27 a number of wave crests are indicated by transverse lines.

16.9 Description of Light Propagation by Travelling Waves. Diffraction

The propagation of waves can be *transversely* bounded by an obstacle, e.g. the blades of a slit. This transverse bounding of a wave can be represented by straight lines or rays, but only to a more-or-less good *approximation*. For this approximation to be valid, two conditions must be fulfilled: The geometric dimensions of the obstacle, e.g. the width B of the slit in Fig. 16.2, must be large compared to the wavelength; and secondly, the point of observation should not be too far from the obstacle. In reality, the geometrically-constructed boundaries of the beam are always exceeded; the waves "run over" their boundaries. This behavior of waves is absurdly described in the *passive form*; we say: The waves are diffracted. Diffraction is inseparably intertwined with every attempt to limit the boundaries of a beam.[C16.18] Figure 16.28 shows a model experiment and reminds us briefly of the phenomenon of diffraction by a narrow slit. According to Vol. 1 (Sect. 12.13), the angular spacing of the first diffraction minimum from the optical axis is given by the equation

$$\sin\alpha_1 = \frac{\lambda}{B}\,, \tag{16.23}$$

and for the angular spacing of the first *sub*-maximum, we found

$$\sin\alpha_1' = \frac{3}{2}\frac{\lambda}{B}\,. \tag{16.24}$$

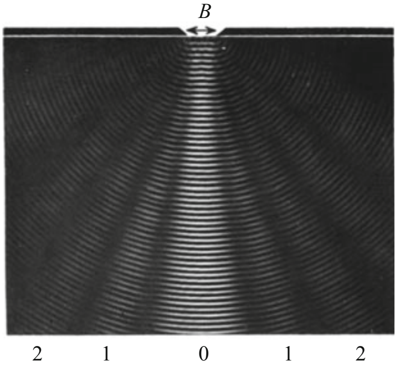
By measuring these angles and the width B of the slit, we can obtain a rather precise determination of the wavelength λ.

These facts as we already know them from Vol. 1 are observed as well for the propagation of light waves. Thus, light waves *cannot* be "cut out" into an arbitrarily narrow beam by the two blades of a slit. Light also exceeds the geometrically-constructed boundaries of a beam; it is "diffracted". In the region of diffraction, we can observe a periodic distribution of the radiation intensity, with maxima and minima.

For a demonstration, we use the setup sketched in Fig. 16.29, top. Compare it with the model experiment illustrated by Fig. 16.28. Note

[C16.18] *Diffraction* was described in detail in Vol. 1 (Chap. 12). Here, we are concerned with showing that this phenomenon also occurs with light waves. A more exhaustive treatment will follow in Chap. 21.

Figure 16.28 A model experiment on diffraction by a narrow slit (compare Vol. 1, Sect. 12.12)



also the scale shown in the legend of Fig. 16.29. Slit *II* is intended to *limit* the light beam to a narrow width, and it should, according to the geometrical construction, give a bright spot on the screen with a width of around 2 mm. Instead, we find the pattern on the screen that is shown as a photograph in Fig. 16.29 (center).

With light, sharply-bounded beams and their representation by means of straight-line rays or lines on the blackboard are also merely an approximation. However, this approximation is often particularly good in optics because of the small wavelengths of the light waves.

In order to obtain a quantitative description of our observations, we measure the distribution of the *irradiance* (i.e. the quotient of radiant power/irradiated area) in our first experiment on diffraction, shown in Fig. 16.29 (see Comment C19.4). We place a narrow slit as aperture in front of our radiometer, so that we use only a strip about 0.5 mm wide of its detector area. Then we put the radiometer at the position of the screen in the light beam and slowly slide it in a direction transverse to the optical axis. At each position, the deflection of the radiometer is noted and then plotted as a graph. In this way, we obtain the curve shown at the bottom of Fig. 16.29; it provides a quantitative complement to the photograph of the diffraction pattern in the center part of the figure.

Making use of Eq. (16.23) and the dimensions given in the figure caption, we determine the wavelength of our red filter light to be about 650 nm. It is around 20 000 times smaller than the wavelengths of the sound waves and water surface waves that we used in mechanics for demonstration experiments ($\lambda \approx 1.2$ cm).

The irradiance plotted on the ordinate is proportional to the power transported by the light waves, and this, in turn, is proportional for all types of radiation to the square of the wave amplitude (see Vol. 1, Sect. 12.24). Therefore, we can state that: The *amplitude of a light wave* is in the first approximation a quantity which is proportional to the square root of the deflection of a radiometer. This may not satisfy our need for concreteness and clarity, but it is sufficient for a quantitative treatment of numerous optical phenomena.[C16.19]

C16.19. See also Comment C12.5.

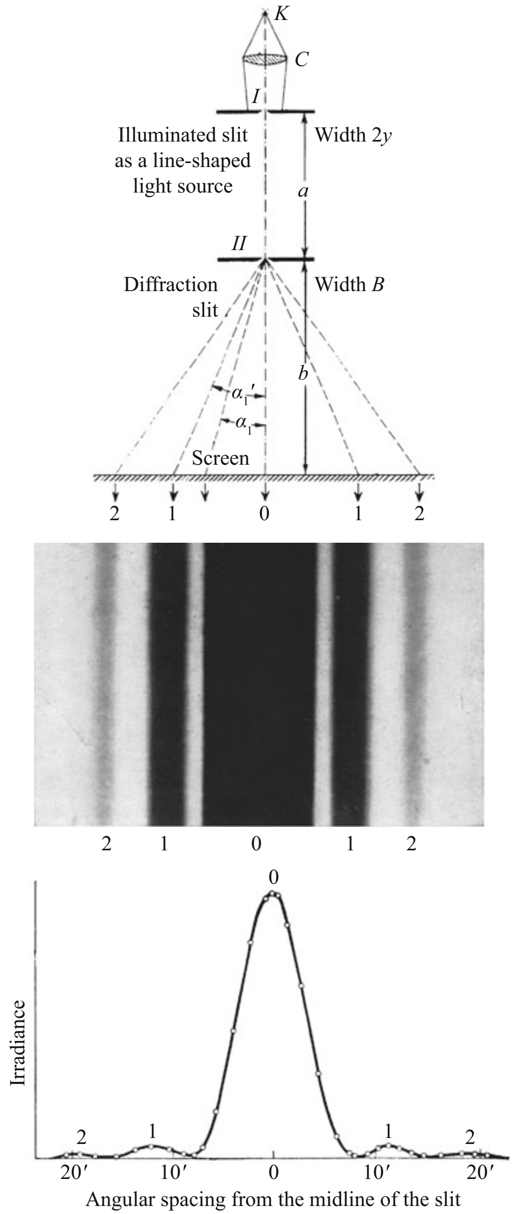


Figure 16.29 Bounding of a beam of light (red filter light) by a slit. *Top part*: The experimental setup. The angles between the dashed rays are greatly exaggerated. *Center*: A short section of the diffraction pattern as observed on the screen (a photographic negative shown actual size for $B = 0.3\,\text{mm}$, $b = 3.8\,\text{m}$, $a = 1\,\text{m}$, $2y = 0.2\,\text{mm}$). *Bottom*: The distribution of the irradiance in the diffraction pattern of a slit as measured with a photodiode (i.e. measured as the quotient of radiant power/area, e.g. in W/m^2). This is the diffraction pattern of a slit ($B = 0.31\,\text{mm}$, $b = 1\,\text{m}$, $a = 0.75\,\text{m}$, $2y = 0.26\,\text{mm}$; the width of the radiometer aperture used here (Fig. 15.6, right) was 0.55 mm) (**Video 16.3**).

Video 16.3:
"Diffraction and coherence"
http://tiny.cc/xdggoy
Beginning at 9:30 min., the **diffraction pattern of a single slit** is demonstrated. Both red and blue filter light beams are used. The experimental setup is explained at the beginning of the video.

16.10 Radiation of Different Wavelengths. Dispersion

We now repeat in Fig. 16.30 the fundamental experiment on refraction shown in Fig. 16.4, however with two modifications. First, we use ordinary natural light (from an incandescent light bulb) instead of red filter light. We allow a narrow, nearly collimated light beam (width < 1 mm) from the aperture B to be incident on the *plane-parallel* glass block. Second, we observe the refracted beam *after* it has emerged from the lower surface of the block, which is exactly parallel to its upper surface. We find an important new result: The beam of light from the incandescent lamp spreads out within the glass block into a series of sub-beams. Parallel, colored light beams emerge from the lower surface of the block. Figure 16.30 shows only a red and a blue sub-beam. In reality, we see a continuous band below the glass block with a whole series of bright colors, i.e. a *continuous spectrum*. We can make this spectrum visible to a large audience; to this end, we need only use a lens l_2 to project the exit position b of the light beam with sufficient magnification onto a screen.

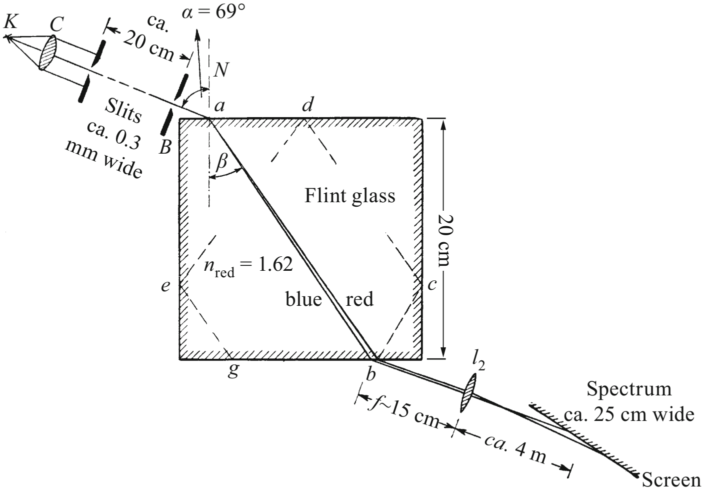


Figure 16.30 Production of a spectrum by refraction in a plane-parallel glass block. The line-shaped light source here and in Fig. 16.31 is a narrow slit S illuminated by an arc lamp K through a condenser lens C. From a onwards to the right and out to lens l_2, the dashes do not refer as usual to rays, but rather to a diverging light beam. The screen must be set up at a grazing angle, so that the chromatic aberration (Sect. 18.8) of lens l_2 is compensated and the band of the spectrum has a practically uniform width. At b, the spectrum is around 2.5 mm wide. However, light reflected along the paths c, d, e, g can be observed at g. There, the spectrum is about 8 mm wide owing to the longer path length of the light within the glass (about a factor of three), and, with lens l_2, it can be projected onto the screen where it has a width of around 75 cm.

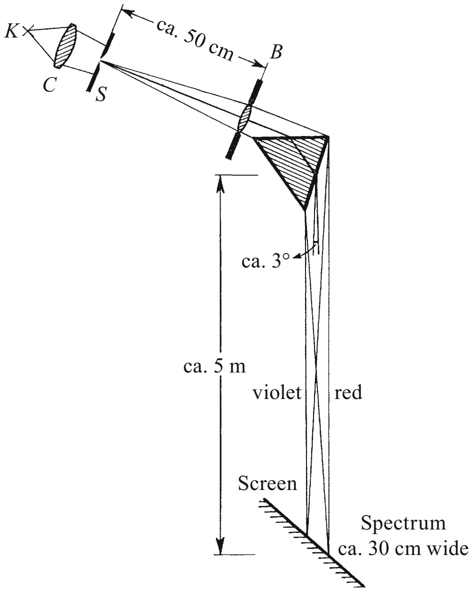


Figure 16.31 Production of a spectrum with a prism, as a demonstration experiment. The light beam which is incident on the prism is only approximately parallel. The lens images the slit (line-shaped light source) onto a wall screen at a distance of several meters. Here, out of the "colored light beam", only the red and violet sub-beams are drawn. The grazing angle of the screen is again necessary for the reason given in the caption of Fig. 16.30. A typical setup for accurate measurements using a precisely collimated light beam is shown later in Fig. 22.2.

Refraction in a plane-parallel glass block thus *generates* a series of colored beams from a beam of colorless natural light. These colored beams are fanned out within the glass block, but after emerging from the block, they are parallel to each other. As previously, we will accept the somewhat vague expression "colored beams" and will first try to continue the fanning-out of the sub-beams outside the block. This can be readily accomplished: We simply have to abandon the parallel form of the upper and lower glass surfaces and instead make the block in the shape of a prism.

With the stronger fanning-out, we can use a much wider light beam than in the case of the plane-parallel glass block. At first, however, we are bothered by the overlap of the individual colored sub-beams; therefore, we make use of the trick explained in Sect. 16.8. We place a lens within the wide aperture *B* and make all the emerging light beams convergent, i.e. we form an image of the illuminated slit *S* on a screen (Fig. 16.31). There, we can clearly see the brightly-colored band of a continuous spectrum.

Now we carry out a quantitative evaluation of these observations. First of all, we need to eliminate the unphysical terms "red", "blue" etc. light beams and instead characterize the various radiations of the

Table 16.2 Indices of refraction at different wavelengths

	Index of refraction at the wavelength			
Material	λ = 656 nm	λ = 578 nm	λ = 436 nm	λ = 405 nm
Light crown glass (BK1)*	1.5076	1.5101	1.5200	1.5236
Light flint glass (F1)*	1.6150	1.6200	1.6421	1.6507
Heavy flint glass (SF4)*	1.7473	1.7552	1.7913	1.8060
Diamond	2.4099	2.4175	2.4499	2.4621

* These are the technical designations for the glass types (Schott AG, Mainz, Germany)

sub-beams physically, i.e. by means of a measurable quantity. This is the concept of the *wavelength*: We separate a narrow light beam from the spectrum, which appears to the eye to contain a single color (i.e. to be "monochromatic"), and, using the now well-known procedure of diffraction by a slit, we measure its wavelength (Fig. 16.29, practical lab experiment). We thus find for a light beam

in the violet	spectral region, wavelengths of 400–440 nm[4]
in the blue	spectral region, wavelengths of 440–495 nm
in the green	spectral region, wavelengths of 495–580 nm
in the yellow + orange	spectral regions, wavelengths of 580–640 nm
in the red	spectral region, wavelengths of 640–750 nm.

The process of refraction thus *produces* a variety of radiations from the radiation of (colorless) natural light, which are colored to the eye, and each of them can be attributed to a *wavelength range* between 400 and 800 nm. In the coming sections, it will suffice to quote an average wavelength for each of these regions. With this, however, we mean a whole range. The same is true of the red filter light which we often employ.

For each such radiation or spectral range, characterized by an (average) wavelength, we can determine separately the index of refraction n of a particular material. In principle, the setup shown in Fig. 16.4 is adequate for this purpose. Then, for various materials which are often used in optics, we obtain the indices of refraction given in Table 16.2.

The dependence of the index of refraction n on the wavelength is called *dispersion*.[C16.20] The technical details of the measurements are unimportant to us. Here, we are first of all concerned with an additional observation which is of fundamental importance. We replace the eye by a physical indicator, e.g. a thermocouple (Fig. 15.5). It can be moved in Fig. 16.31 along the band of the spectrum so as to measure the radiant power as a function of wavelength. The deflection of the indicator does not go to zero beyond the ends of the

C16.20. The dispersion which occurs in the range of visible light, in which the index of refraction decreases with increasing wavelength, is termed *normal* dispersion. In other wavelength ranges, the relation between wavelength and index of refraction is more complex; see Fig. 27.1 (bottom right) and 27.2.

[4] 1 nm = 1 nanometer = 10^{-9} m.

visible spectrum, that is beyond the border of the violet at one end and the red at the other. On the contrary, on both sides of the visible spectrum, we observe a considerable signal from invisible radiation. Refraction thus produces not only a visible spectrum, but also invisible light beams outside the visible range. These regions are given the collective names “ultraviolet”[5] and “infrared”[6].

For our earlier demonstration experiments, we produced red light not by refraction, but instead by employing a red *filter*: We passed the natural light from an arc lamp through a piece of red glass. The word filter is based on an old custom: One describes the colorless radiation of an incandescent lamp as a *mixture*[7] of various colored radiations having different wavelengths. The ‘filter’ allows only one of them to pass through.

In a corresponding manner, we can also prepare filters for the invisible radiations. As a filter for the ultraviolet, we can most conveniently use a piece of glass with a high content of nickel. To the eye, it appears completely black and opaque; but in the language of the above description, it allows ultraviolet light from the mixture of radiation emitted by the arc lamp to pass through. In order to make the ultraviolet light beam visible, for demonstration experiments we can use excitation of fluorescent radiation. Numerous substances “fluoresce”, i.e. they glow brightly with visible light when they are struck by ultraviolet light. Thus, in Fig. 16.3, instead of a red filter, we make use of an ultraviolet filter and paint the board which serves as screen with a fluorescent pigment, e.g. a layer of paint containing a powdered zinc salt. A bright, greenish fluorescence indicates the trace of the invisible ultraviolet collimated light beam.

As a filter for infrared radiation, glass plates containing manganese oxide (MnO) are suitable. To detect the infrared radiation, one usually employs the warming of the irradiated object. Thus, in Fig. 16.32, we set up a searchlight with infrared light by using an arc lamp and an infrared filter, and we can ignite matches at a distance of 10 m using its invisible radiation.

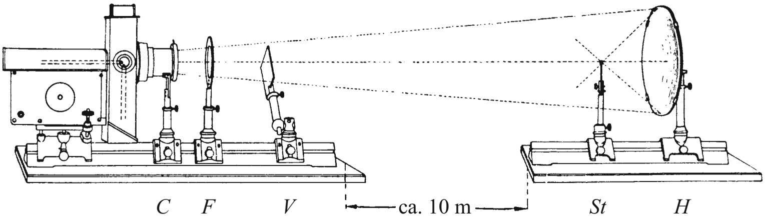


Figure 16.32 Igniting a match *St* by means of a beam of invisible infrared radiation (*C* is an auxiliary condenser lens, *F* is an infrared filter, *V* is the shutter and *H* is a concave mirror)

[5] J.W. Ritter: Gilberts Ann. **7**, 527 (1801).

[6] F.W. Herschel: Philosophical Transactions, Part II, 284. London 1800.

[7] This old-fashioned terminology is often convenient, but according to the current state of knowledge, it is at best rather imprecise (see Sects. 20.11 and 22.4).

16.11 Some Technical Resources: Angled Mirrors and Mirror Prisms

Angled mirrors and mirror prisms are frequently-used technical aids to experimentation. In addition, the role of refraction and dispersion in mirror prisms provides us with a useful subject for contemplation.

Often, one wants to deflect a light beam by a certain angle δ. This can be most simply accomplished by a single mirror reflection according to the scheme shown in Fig. 16.33. But this setup is overly sensitive to a sideways tilting of the mirror; tilting through an angle σ (tilt axis perpendicular to the plane of the page) changes the angle δ between the incident and the reflected beam by 2σ.

If we use a double reflection by an angled mirror, in contrast, sideways tipping of the mirror has no effect, since, as can be seen in Fig. 16.34, the angle δ between the incident and the doubly-reflected

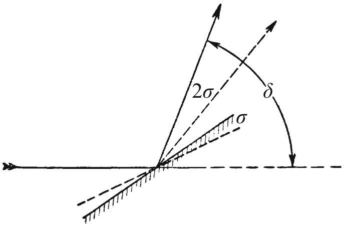


Figure 16.33 The influence of tilting a mirror on the direction of a reflected light beam (only the central axis of the beam is shown)

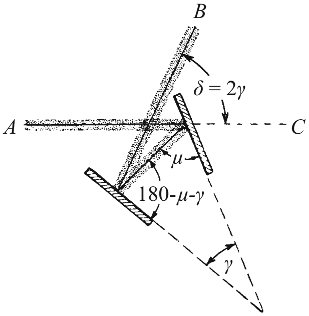


Figure 16.34 An angled mirror with a measurable wedge angle γ permits a free-hand measurement of the angular spacing δ between two objects in the directions B and C (the sextant used by seamen and astronomers). Imagine that your eye is at A and that the right face of the mirror is partially transparent, e.g. only half silvered.

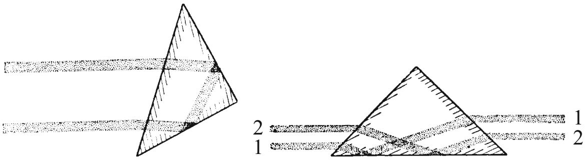


Figure 16.35 Mirror prisms with right-angled triangular shapes. At the left: with silvered cathetus faces for reversing beam directions; at the right: with a silvered hypotenuse face for exchanging light beams 1 and 2. A *reversing prism* can be used to invert images which are upside down, e.g. for the projection of physics demonstration experiments (**Exercise 16.4**).

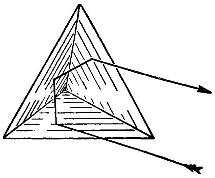


Figure 16.36 The optical path through a right-angled corner mirror: The principle of the "cat's-eye mosaic"

beam depends only on the wedge angle γ between the two mirror faces. We find

$$\delta = 2\gamma\,. \tag{16.25}$$

For a beam deflection of 90°, γ must be chosen to be 45°. Two mirror faces which are perpendicular to each other ($\gamma = 90°$) yield $\delta = 180°$, that is they reflect the incident beam backwards, parallel but opposite to its original direction, etc. (Fig. 16.35). The same effect is produced by a corner mirror (Fig. 16.36).

Exercises

16.1 a) Derive Eq. (16.8) for a prism by application of the law of refraction (16.2). Note: Apply the law of refraction to each face of the prism, take the sum and the difference of the resulting expressions and rewrite them as products. b) A light beam is incident at an angle $\alpha_1 = 30°$ onto a 60°-prism with an index of refraction of $n = 1.5$. Compute the angle of deflection δ, by which the light beam is refracted in total. Note: Calculate β_1 from Eq. (16.2) and insert its value into Eq. (16.8). (Sect. 16.6)

16.2 Show that for image formation using a concave mirror (Fig. 16.24), for near-axial light beams, the focal length f is equal to half the radius of curvature R of the mirror. (Sect. 16.7)

16.3 Show that the lens formula for near-axial light beams (Eq. (16.13)) is also valid for imaging by a concave mirror. (Sect. 16.7, 18.2)

16.4 In mirror prisms (Fig. 16.35), in addition to reflection, refraction also always occurs. If natural light (from an incandescent lamp) is used, one thus also observes dispersion (Sect. 16.10). After the light beam emerges from the prism, the different sub-beams, which are refracted through different angles, are shifted parallel to each other. Why can one still see the object without colored fringes when using a mirror prism? (Sect. 16.11)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_16) contains supplementary material, which is available to authorized users.

Image Formation and Light-Beam Boundaries 17

17.1 Fundamentals of Image Formation

For image formation, basically only one condition needs to be fulfilled: The rays which originate at "object points" must be limited to a small spatial *opening angle ω* by passing through an *aperture* (Sect. 16.1) on their way to the image plane (e.g. a projection screen or a photographic plate). In Fig. 17.1a, the aperture *B* is a small round opening (pinhole camera!); in Fig. 17.1b, it is a small planar mirror. In both cases, as a suitable *object* to be imaged, we have chosen an arrow or letter made with small light bulbs.

Which opening angle *ω* for the radiation gives the sharpest "image points" (or, in modern terms, "pixels") for a given distance between the object and the image plane? (The image points are small circular disks which make up the image.) The answer will be given later in Sect. 21.8. The decisive quantity which determines it is the *wavelength*, which is a property of the radiation.

Before answering this question, we turn to image formation using lenses and concave mirrors, due to their practical importance. This is nothing fundamentally new; it simply makes it possible to use larger opening angles *ω* and thus to permit the image to be better resolved. A lens in, in front of, or behind the aperture makes the *optical path length* (Sect. 16.3, see also Fig. 16.18) for all the rays between the object and the image points, even within a large opening angle, practically the same; e.g. for the rays 1 and 2 in Fig. 17.1c. A concave

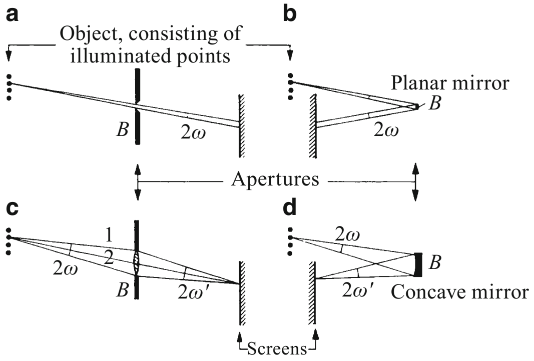


Figure 17.1 The role played by the aperture *B* in image formation

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_17

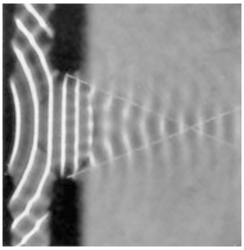


Figure 17.2 A shallow-water lens for surface waves on water. It images an "object point", the wave center located to the left of the figure on the optical axis, as an "image point" in the image plane. The image point has a finite extension, as a result of diffraction by the opening of the lens mount. (See Vol. 1, **Video 12.2**)

Video 12.2 (Vol. 1): **"Experiments with water waves"** http://tiny.cc/tfgvjy.

mirror has the same effect owing to its curvature (Fig. 17.1d). This is all that we can say within the geometrical-optical picture which describes the light beam in terms of rays. We can penetrate further into the subject of image formation only by considering the wave nature of the radiation (i.e. its wavelength). This is shown in detail in Vol. 1, in particular in Figs. 12.20, 12.21 and 12.31 in Chap. 12 of that volume. We will refer to those figures in the following section.

17.2 Image Points as Diffraction Patterns from the Lens Mount

With Fig. 17.2, we remind the reader of an important result which we obtained by considering mechanical waves: An image point from a lens is *not* the intersection of two geometric straight lines (rays), but rather a diffraction pattern[C17.1] from the mount (i.e. the outer boundary) of the lens or mirror. This diffraction pattern has a finite extension; it is not a "geometrical point".

C17.1. The diffraction of light was introduced already in Sect. 16.9. A more detailed treatment will follow in Chap. 21.

For light waves, this important fact can be demonstrated with the apparatus sketched in Fig. 17.3. There, we form the image of a lattice of points on a screen located 5 m away, using a good-quality telescope lens L_1 (an objective with a focal length of 70 cm). The point lattice (3 mm on a side) is composed of 25 object points, each one a round hole of 0.2 mm diameter, and is illuminated from behind by intense red light. The light beam which emerges from the lens is bounded by

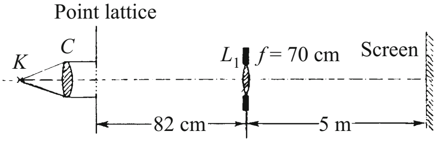


Figure 17.3 Imaging of a small square point lattice using a telescope objective. The lattice consists of 25 holes, each 0.2 mm in diameter, at a spacing of 0.7 mm; see Fig. 17.4. For demonstration in a large room, f (the focal length of the lens) must be chosen to be shorter.

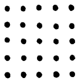


Figure 17.4 The point lattice as imaged on the screen in Fig. 17.3 (photographic negative, 1/2 actual size)

the circular lens mount (diameter 5 cm)[1]. The image on the screen has been photographed in Fig. 17.4. The photo shows a lattice composed of 25 cleanly separated circular disks. They give an *upper limit* for the diameter of an "image point". Now, we place an ancillary aperture B_1 just in front of the object (Fig. 17.5), and let only the central hole be imaged, that is one single "object point". Its image remains on the screen, maintaining its sharpness.

Now the decisive observation: We add a second aperture in the form of a rectangular slit B_2 just behind the lens in Fig. 17.5. This produces a rectangular boundary for the light beam emerging from the lens, for example with a width of $B = 0.3$ mm. On the screen, we see the image which is reproduced as a photo in Fig. 17.6 (1/2 actual size): The object point is imaged in the image plane as a long "brushstroke", with shorter replicas on each side. Using blue-filter light, we see the same "brushstrokes", but they are somewhat shorter (Fig. 17.7).

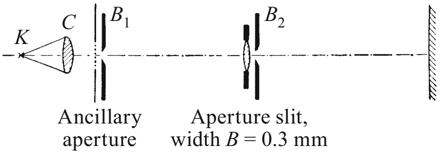


Figure 17.5 An ancillary aperture B_1 covers 24 of the 25 openings in the point lattice (object). The remaining opening is imaged by the same objective lens as in Fig. 17.3; but this time, the light beam is bounded with a rectangular shape by a second aperture B_2.

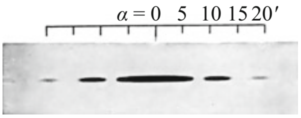


Figure 17.6 The image point formed by a lens whose beam has been limited by a narrow rectangular slit B_2, 0.30 mm wide and perpendicular to the long direction of this figure; it has the form of a brushstroke. The image was photographed using red-filter light ($\lambda \approx 660$ nm) at a distance of 5 m. (Photographic negative, 1/2 actual size, $\alpha = 20'$ corresponds to $\sin\alpha = 5.8 \cdot 10^{-3}$ rad). For angular units, see the footnote at the end of this chapter.

[1] The illumination lens or condenser C has to form an image of the crater K of the arc lamp (the source point of the illumination) in the plane of L_1, and its diameter must be greater than that of the lens L_1.

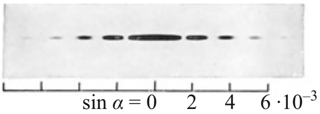


Figure 17.7 As in Fig. 17.6, but using blue-filter light, with $\lambda \approx$ 470 nm

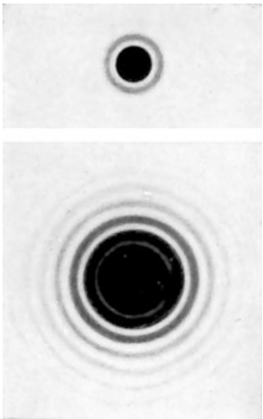


Figure 17.8 The image point cast by a telescope objective lens through a circular aperture of 1.5 mm diameter, photographed at a distance of 5 m (top picture 1 min., bottom picture 5 min. exposure time; red-filter light, actual-size negatives)

In both cases, the images resemble a horizontal section from the well-known diffraction pattern cast by a slit (Fig. 16.29, center). The minima are at the same spacing as before (compare Fig. 17.6 with Fig. 16.29, bottom). Therefore, we can explain Figs. 17.6 and 17.7 with some certainty: *An image point is in reality a diffraction pattern from the boundary of the lens*. Its first minimum appears, as seen from the center of the lens, on both sides of the midpoint of the image at an angle α, as defined by Eq. (16.23)[2]

$$\sin\alpha = \frac{\lambda}{B}\,. \qquad (16.23)$$

Normally, the boundary or mount of a lens is not rectangular, but rather circular; then instead of a slit, we have the circular opening of the lens mount. Therefore, in continuing our experiments, we replace the slit B_2 in Fig. 17.5 by a circular aperture (e.g. with a diameter of 1.5 mm). The result can be seen in Fig. 17.8; it is, as expected, the diffraction pattern of a circular aperture. Qualitatively, we can say that it is formed by rotation of the diffraction pattern from a slit around its center point (Fig. 17.6). Quantitatively, this is not quite correct; in the case of a circular opening, we have to add a numerical factor of 1.22 on the right-hand side of Eq. (16.23).[C17.2] This is, however, practically unimportant, given the wide range of wavelengths λ in the visible spectral region (around 400–750 nm).

C17.2. For the calculation of the value of this factor, 1.22, see e.g. Max Born and Emil Wolf, *Principles of Optics* (Pergamon Press, 4th edition (1970)), Sect. 49 (available online – see Comment C25.11).

Result: *The image point cast by a lens is a diffraction pattern of the opening which limits the effective lens diameter*. We can state

[2] In Fig. 17.5, we can consider the light beam which falls on the slit B_2 to be a collimated beam to a good approximation. Thus the condition for the validity of Eq. (16.23) is fulfilled.

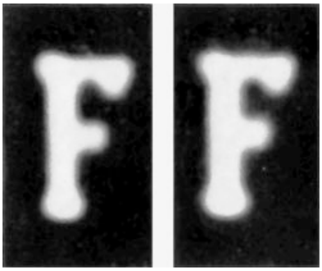


Figure 17.9 Photographs of two images of the same letter 'F' (a copper stencil), formed on the left using a lens, and on the right using a pinhole camera. The lens mount and the aperture of the pinhole camera are both 3.5 mm in diameter (!) Object and image distances are each 17 m. The images were thus just the same size as the object. They are reproduced here actual size. The surprisingly large "image points" which compose the images are shown later in Fig. 21.19.

without serious exaggeration that in imaging by lenses, the opening which bounds the light beam is more important than the lens itself. The role played by the lens is only secondary; it makes the incident plane or divergent wave train convergent and focusses it into a small region. In this way, it places the diffraction pattern from its opening at a conveniently accessible distance. The image which is composed of these "diffraction-pattern image points" is then formed at a manageable size.

"... in imaging by lenses, the opening which bounds the light beam is more important than the lens itself".

If we remove the lens, the remaining aperture acts just like a pinhole camera (Fig. 18.35). In the latter, the diameter of the aperture must be adjusted to suit the desired distances of the object and the image plane. This topic, as mentioned at the beginning of this chapter, will be treated in Sect. 21.8. If this adjustment has been carried out correctly, the resulting sharpness (resolution) of the image cannot be further improved by inserting a lens into the aperture. This is demonstrated in Fig. 17.9 for an aperture of 3.5 mm diameter.

There is no fundamental difference between the image points of a pinhole camera and those of a lens. Both are merely diffraction patterns of the aperture. To be continued in Sect. 21.8...

17.3 The Resolving Power of Lenses, Particularly in the Eye and in Astronomical Telescopes

The significance of the experiments just described will be explained by giving some examples. We return to Fig. 17.5, remove the ancillary aperture B_1, and thereby form an image of all 25 object points in the point lattice. Then we again add a rectangular aperture to the

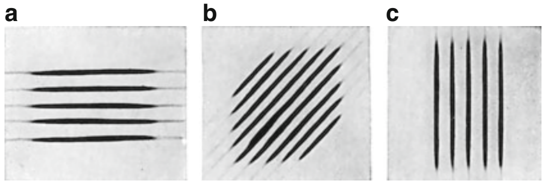


Figure 17.10 The images of the point lattice in Fig. 17.3 are drastically modified, depending on the shape and size of the boundary of the objective (red-filter light, photographic negative, 1/2 actual size)

lens, that is we again produce long "brushstrokes" as image points (Fig. 17.6), initially with a horizontal orientation (the slit B_2 is vertical). Then the lens forms the image shown in Fig. 17.10 a. Instead of the point lattice (Fig. 17.4), we see an image of five bright horizontal lines which result from the overlap of the horizontal brushstrokes. We then rotate slit B_2, and with it the brushstrokes, by an angle of 45° to the vertical. Now, instead of a point lattice, we see the image as shown in Fig. 17.10b, etc. One inappropriate restriction of the light beam can thus make the image completely different from the object it is supposed to represent.

With the usual shape of the lens boundary, a circular lens mount, we obtain circular diffraction disks as image points, surrounded by concentric rings of decreasing intensity (Fig. 17.8). Seen from the center of the lens, the first ring of minimal intensity (which appears bright in the photographic negative) in this diffraction pattern is at an angular spacing of α from the center of the diffraction disk. Then for a lens diameter B, we find to a good approximation

$$\sin\alpha = \frac{\lambda}{B} \quad \text{or} \quad \alpha \approx \frac{\lambda}{B}\,. \tag{16.23}$$

In order to separate (resolve) two object points in the image, they have to be roughly as far apart as shown in Fig. 17.11: The central disk of the one image point must fall in the first minimum of the other (or be still further away).[C17.3] That is, the angular spacing 2φ of the object points should not be much smaller than the angle α calculated from Eq. (16.23). Then we find for the smallest "resolvable" angular spacing

$$2\varphi_{\min} \approx \frac{\lambda}{B}\,. \tag{17.1}$$

Example: Our eyes are basically similar to photographic cameras. Instead of the photographic film or plate (or CCD detector array), they contain the *retina*, which takes the form of a mosaic. The boundary of the lens of the eye ($f = 23\,\text{mm}$) is defined by the iris (pupil). Its opening diameter in daylight is about 3 mm. Taking $\lambda = 600\,\text{nm} = 6 \cdot 10^{-4}\,\text{mm}$ as the average wavelength of sunlight, we obtain from

[C17.3] C17.3. This condition is known as the "Rayleigh criterion" in the literature.

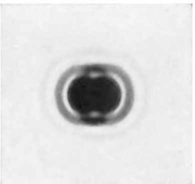


Figure 17.11 The resolving power of a lens: Separating the two diffraction patterns (from a circular lens aperture of 1.5 mm diameter) which appear as neighboring image points. The object consisted of two holes of 0.2 mm diameter at a spacing of 0.3 mm (photo taken with red-filter light at a distance of 5 m; photographic negative, actual size) **(Video 17.1)** .

Eq. (17.1)[3]

$$2\varphi_{\min} = \frac{6 \cdot 10^{-4}\,\text{mm}}{3\,\text{mm}} = 2 \cdot 10^{-4} = 41 \text{ arc second}\,.$$

This means that our eyes must just be able to distinguish two object points at an angular spacing of around 1 arc minute; or, in other words: About 1 arc minute is the smallest "resolvable" angle of sight 2φ for the eye (cf. Fig. 18.23). This rough estimate agrees with practical experience. For a demonstration, we can use a black-and-white square lattice. For an observer at a distance of 10 m, the spacing of the lines on the lattice has to be about 3 mm in order to resolve them visually. It then follows that:

$$2\varphi_{\min} = 3 \cdot 10^{-4}$$

or

$$2\varphi_{\min} = 1 \text{ arc minute}\,.$$

With optimized illumination, we can approach roughly half this value. We thus do not have to make the spacing quite as great as shown in Fig. 17.11.

The *astronomical telescope* is also practically just a variant of a photographic camera: A lens or a concave mirror, with a photographic plate or photodetector in its focal plane. For a lens or mirror diameter of 300 mm, the smallest resolvable angle of vision is 100 times smaller than for the naked eye, i.e. about 0.4 arc seconds. With a diameter of 1.2 m, we can still resolve two fixed stars at a spacing of 0.1 arc second, etc. Each of the two stars is visible only as a diffraction disk from the lens or mirror mount. If the objective of a telescope has a triangular boundary, the diffraction pattern from a fixed star looks like that shown in Fig. 17.12. A true image of the disks of fixed stars, corresponding to an image of the Sun's disk, cannot be obtained

Video 17.1:
"Resolving power"
http://tiny.cc/9dggoy
The video shows that "image points" are simply diffraction patterns from the lens mount. The resolving power which is limited by the diameter of these diffraction patterns is demonstrated using two neighboring circular openings at different spacings, illuminated with red and with blue light. A preliminary experiment using a single opening demonstrates the reduction of the effective lens diameter using a circular aperture. The initially sharp image is gradually converted into the enlarged pattern of a diffraction disk as the lens diameter is decreased (see Sect. 21.3).

[3] Units: 1 degree (°) = $1.745 \cdot 10^{-2}$ rad, 1 minute of arc (′) = $2.91 \cdot 10^{-4}$ rad, 1 second of arc (″) = $4.86 \cdot 10^{-6}$ rad (cf. Vol. 1, Sect. 1.5).

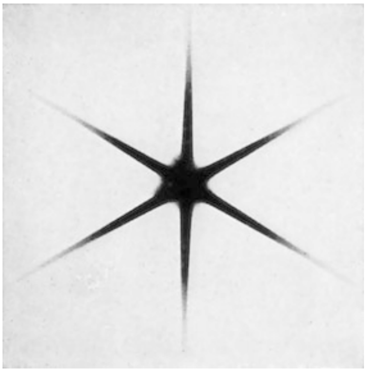


Figure 17.12 The image point formed by a lens which is bounded by a triangular aperture with a side length of 9 cm (at a distance of 5 m, taken with red-filter light, negative shown actual size)

with earthbound telescopes. The diameter of the Sun's disk is 32 arc minute, while the diameters of the disks of even nearby stars is less than 0.01 arc second. To form images of the disks of fixed stars, the diffraction-limited image points of even the largest telescopes (mirror diameter for example 5 m)[C17.4] are still much too large.

The resolving power of the eye and of telescopes is determined by the *boundary of the light beam*, not by details of the lens construction. That is the important result of this section.

C17.4. Here, Pohl is referring to the *Hale Telescope* operated by the California Institute of Technology (Caltech) since 1948 on Mount Palomar in the neighborhood of San Diego, CA/USA. A considerable perturbation is due to the Earth's atmosphere. For example, using the HUBBLE Space Telescope, which has a mirror diameter of only 2.4 m at an altitude of 500 km – that is, far outside the perturbing influence of the atmosphere, and which orbits the earth as an artificial satellite, it is possible to observe the surface of the star α Orionis (Betelgeuse). A large, bright spot was seen there. The angular diameter of the disk amounts to 0.047 arc second. It was determined using FIZEAU's method (Sect. 20.16) (See e.g. R.L. Gilliland and A.K. Dupree, *Astrophysical Journal* **463**, L39 (1996)) (**Exercise 17.2**).

Exercises

17.1 The International Space Station (ISS) circles the Earth at an average altitude of ca. 350 km.
a) What is the minimum spacing d of two lighthouse beams so that they can be seen as separate (resolved) by an astronaut with the naked eye?
b) What diameter B would a camera objective require at this altitude in order to resolve two persons on the ground standing 2 m apart (corresponding to the RAYLEIGH criterion) in the camera image? (Use the average wavelength of sunlight, 600 nm.) (Sect. 17.3)

17.2 Compare the smallest resolvable angular spacing $2\varphi_{\min}$ for the HUBBLE Space Telescope at a wavelength of $\lambda = 250\,\text{nm}$ with the angle 2α under which the disk of the star Betelgeuse appears from the Earth ($\alpha = 0.047$ arc second). (Sect. 17.3)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_17) contains supplementary material, which is available to authorized users.

Fundamental and Technical Details of Image Formation and Beam Limitation

18

18.1 Preliminary Remark

In optics, *lenses* play roughly the same role as wire leads in electricity and magnetism. Both are indispensable aids to experimental observations. How to use wire leads can be quickly learned and is for the most part compatible with everyday experience. How to make effective use of lenses, in contrast, requires rather extensive, detailed knowledge. The six pages in Sect. 16.7 are by no means sufficient. In particular, the most important aspect is lacking there: The major role played by *light-beam delimitation* in all questions relating to image formation. We met up with that aspect only briefly in Sect. 17.2. In the present chapter, we offer additional examples, once again derived directly from experiments and individual observations.

18.2 Principal Planes, Nodes

In dealing with simple thin lenses, their focal lengths, the distance to the object and the distance to the image from the midplane of the lens are the significant quantities. This midplane is also used in the well-known geometric constructions of image formation which are popular in high-school optics treatments (Fig. 18.1).[C18.1]

We thus neglect the finite thickness of the lenses, considering it to be insignificant. This approximation is, however, nearly always not admissible for thicker lenses and for compound lenses (e.g. objective lenses for microscopy and photography). Then, the description of the optical path of the rays requires more than a single midplane: We must introduce two reference planes perpendicular to the optical axis, the two principal planes H and H'; and we must measure the focal lengths and the object and image distances with reference to these planes (C. F. GAUSS). Likewise, for the geometrical construction of the image, the rays must be extended to one of the principal planes

C18.1. POHL consistently denotes quantities on the *object side* of the lens (as we have already seen in Chap. 16) by letters without a prime, and those referring to the *image side* by letters with a prime, even when they are not distinct, as is the case here and in the following figures. f and f' are unequal when materials with differing indices of refraction are located on the two sides of the lens, as for example with the lens of the eye. See also Fig. 18.24.

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_18

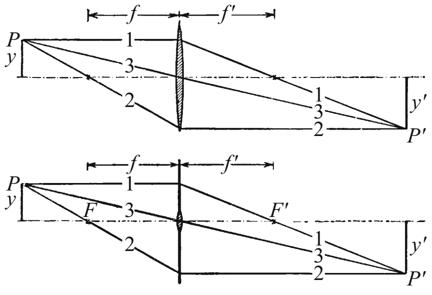


Figure 18.1 The geometric construction of the image point P' which corresponds to the object point P. The focal points F and F' are shown; they are properties of the lens. It suffices to trace any two of the three rays 1–3. *This construction is purely formal.* The object size $2y$ can be arbitrarily greater than the diameter of the lens, e.g. in the case of a photographic camera. Then the rays 1 and 2 themselves no longer intersect the actual lens, but they intersect its midplane. Nevertheless, they are 'bent' at the midplane, as seen in the lower part of the figure (**Exercises 18.1, 16.3**).

and then 'bent' there. This is illustrated in Fig. 18.2. The physical meaning of this construction can be seen in the demonstration experiments shown in Fig. 18.3. The beam axes (rays) which pass through F are called *image-side telecentric*, while those that pass through F' are termed *object-side telecentric*.

Figure 18.2 also illustrates a general definition of the focal lengths, namely

$$\text{on the image side: } f' = \frac{y}{\tan\varphi'}\,, \quad \text{and on the object side: } f = \frac{y'}{\tan\varphi}\,. \tag{18.1}$$

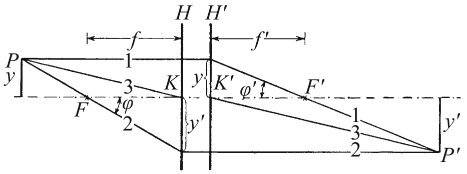


Figure 18.2 The definitions of the object-side and image-side principal planes H and H'. With thick lenses and compound lenses, we measure the object distance on the object side and the image distance on the image side, as well as the corresponding focal lengths. The points K and K' are shown for comparison to Fig. 18.5. If we want to measure the focal length using Eq. (18.1), and therefore want to make e.g. ray 2 the central axis of a light beam, we have to place an aperture around the focal point F as an entrance pupil. This makes ray 2 a principal ray, and thus its angle of inclination to the optical axis on the object side is conventionally denoted as φ.

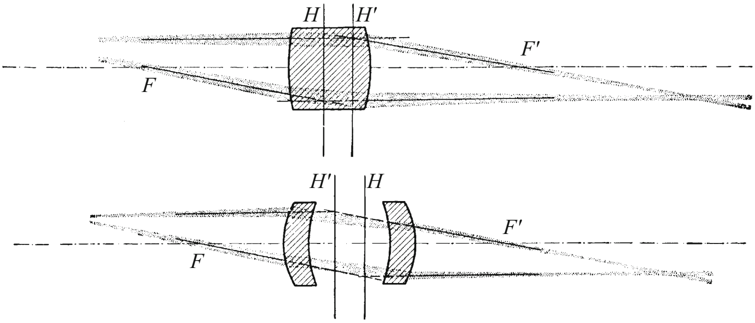


Figure 18.3 Demonstration experiments for illustrating the schematic ray diagram in Fig. 18.2 (red-filter light). For clarity, only the light beams belonging to the rays 1 and 2 are shown (1/9 actual size). In the bottom part of the figure, the image-side principal plane H' is closer to the object than the object-side plane H! (Drawings from photographs of the demonstration experiments. This applies also to later figures, such as Figs. 18.4, 18.10 and 18.13).

To locate the principal planes experimentally, one makes use of two telecentric light beams. They are incident first from the right, parallel to the optical axis (Fig. 18.4, top) and then from the left (Fig. 18.4, bottom). One finds the positions of the focal points F and F' by extrapolating the central rays (short-dashed lines) until they intersect. With this compound lens, the two principal planes H and H' are no longer *between* the individual lenses (a large converging lens and a small diverging lens). Furthermore, we clearly see the very unequal distances of the two focal points from the central plane of the compound lens.

In the usual applications, the same substance, namely air, is found on the object and the image side of the lens. In certain cases, however,

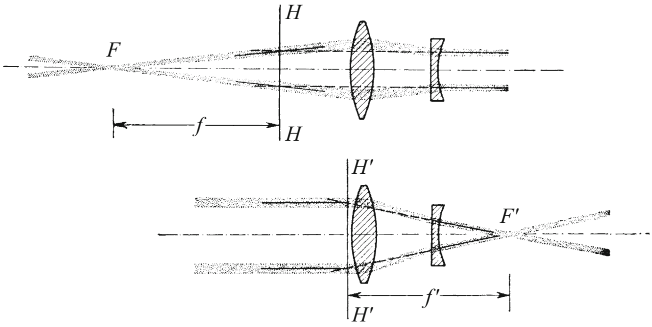


Figure 18.4 A demonstration experiment for locating the principal planes of a compound lens composed of a converging lens and a diverging lens. This type of compound lens is often used in cameras as a *telephoto objective* for taking large close-ups of distant objects, for example wild animals. This requires a long focal length (cf. Eq. (16.22)).[C18.2]

C18.2. With today's telephoto objectives, the focal length f' can be twice as long as the whole lens system. In Fig. 18.4, the principal plane H' would then be at the left margin of the figure.

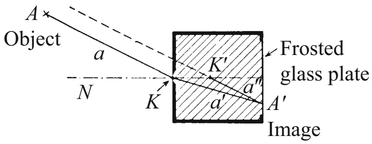


Figure 18.5 The locations of the two nodal points K and K' in a pinhole camera filled with water. The aperture through which the image is formed is kept from leaking by a thin glass plate.

the image side contains a different substance, often a liquid (lens of the eye!). We then require the concept of *nodal points*. It can be most simply explained for the special case of a pinhole camera which is filled with water in its interior (Fig. 18.5). The image formation of the object point A as an image point A' can be described in two ways: Either by referring to the two rays a and a', which are *'bent'* relative to each other by refraction, or by using the rays a and a'', which are parallel to each other on the object and the image sides, and whose intersection points with the optical axis (the dot-dashed symmetry axis of the imaging aperture) define the two points K and K', called the *nodal points*.

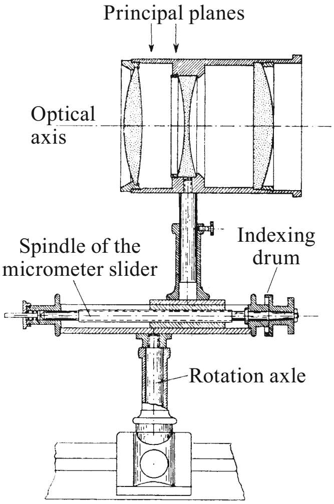
In a corresponding way, we define the nodal points even when a lens is inserted into the aperture. As an example, consider the eye (Fig. 18.24). On the object side, it is bounded by air, and on the image side, the interior of the eyeball; it contains a gelatinous liquid called the *vitreous humor*.

The two nodal points of the relaxed (unaccomodated) eye lie (for normal vision, not peripheral vision) 7.0 and 7.3 mm behind the outermost point (crest) of the cornea (the outer surface of the eye). The principal planes, in contrast, are only about 1.35 and 1.65 mm behind the crest of the cornea, while the focal point lies within the eyeball $(22.8+1.6) = 24.4$ mm behind the crest of the cornea.

In general, however, the same material is found on both sides of the lens. Then the intersection points of the optical axis with the principal planes (the 'principal points') become the "nodal points" K and K': This means that rays which pass through them are parallel to each other on both the object side and the image side. Two such rays (denoted by 3) are drawn in Fig. 18.2.

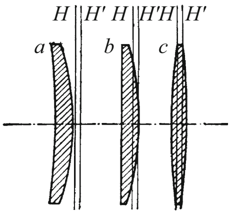
This property of the nodal points can be used to locate the principal planes experimentally. The compound lens is mounted on a slider, with its optical axis (the symmetry axis or lens axis; shown dot-dashed) parallel to the spindle of the slider (Fig. 18.6). The slider is in turn mounted on a vertically-rotatable axle. Then, the lens is used to cast the image of a stationary light source onto a distant screen, and the optical axis is swung back and forth. This in general causes the image on the screen to move; by shifting the slider, we can find a position where this motion ceases. In that case, the axis of rotation is directly under the object-side nodal point, and the axis of rotation lies in the object-side principal plane.

Figure 18.6 Illustration of the experimental determination of the object-side principal plane by locating the object-side nodal point. The lens can be rotated around a vertical axis and can be slid with the slider relative to that axis.



When a higher precision is required, one must determine both of the principal planes, even for simple lenses of moderate thickness; approximating them by the central plane of the lens is not exact. Figures 18.7 and 18.8 show some examples.

Figure 18.7 The principal planes of three thin lenses. Even in the case of the meniscus lens *a*, they deviate only slightly from the central plane of the lens. (A "meniscus lens" is a lens with one concave and one convex curved surface). (1/6 actual size; $f_a = 28\,\text{cm}$, $f_b = 20\,\text{cm}$, $f_c = 21\,\text{cm}$)



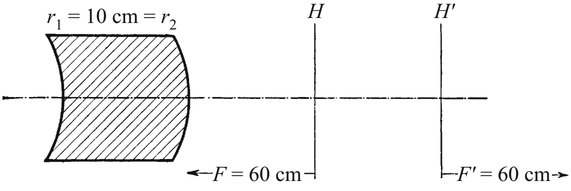


Figure 18.8 A 10 cm-thick meniscus lens, which in spite of having the same curvature on both sides is a *convergent* lens, with its principal planes far outside the lens itself (also shown 1/6 actual size). See the references in Comment C16.14, and also e.g. http://ocw.mit.edu/courses/mechanical-engineering/2-71-optics-spring-2009/video-lectures/lecture-5-thick-lenses-the-composite-lens-the-eye/MIT2_71S09_lec05.pdf

18.3 Pupils and the Boundaries of Light Beams

The contents of this section are especially important. The rays sketched in Fig. 18.1 are possible central axes or delimiting rays of light beams; they are compatible with the positions of the focal points F and F'. But these light beams must by no means actually be present in reality. The light beams which are present in fact often look very different from the rays that we draw in geometric diagrams. Their form is determined by pupils. *A pupil is (on both the object side and the image side) a cross-sectional area common to all the light beams*. It is called the "entrance pupil" for light beams on the object side, and the "exit pupil" for light beams on the image side (ERNST ABBE, 1840–1905).

Examples:

1. In the simplest application of a lens, e.g. as in Fig. 16.21, the lens mount (the "bezel") delimits the light beams coming from the object side (with an opening angle of ω) and acts thus as an "entrance pupil". It also delimits the image-side light beams (opening angle ω') and therefore acts as an "exit pupil". In this simplest example, the two pupils merge into one.

2. In Fig. 18.9, bottom, there is a circular aperture B in front of the lens. It serves as entrance pupil and delimits the object-side light beams (opening angle ω). Behind the lens is its real image B'. This aperture image acts as exit pupil and delimits the image-side light beams (opening angle ω'). Follow the heavy lines from the lower

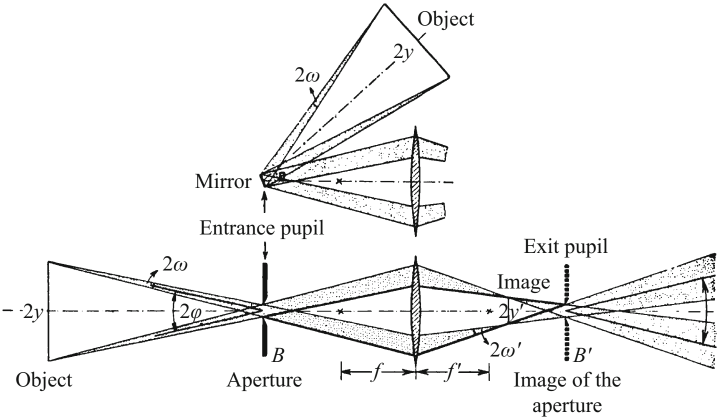


Figure 18.9 Delimiting of the imaging light beams by pupils. The entrance pupil is an aperture in both figures; below, it is a transmission aperture (opening), and above, it is a mirror (reflection aperture). The real image of B acts as exit pupil. φ and φ' are the object- and image-side principal-ray angles of inclination. The lens bezel acts as a delimiting aperture for the field of view (Sect. 18.14).

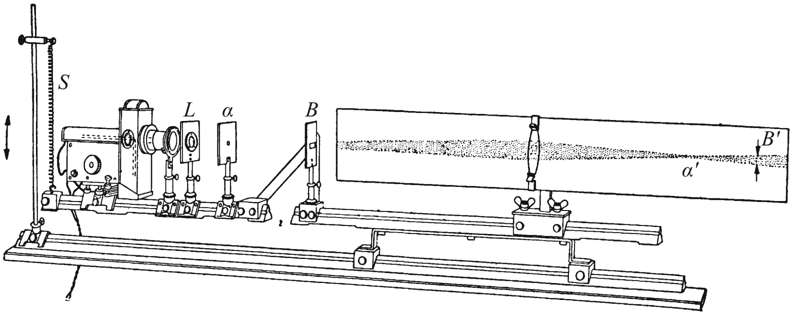


Figure 18.10 A demonstration experiment for showing the locations of pupils. One half of an optical bench can be rotated around the center of the entrance pupil B and is hung from a spring S. Thus, the object point α, an opening which is illuminated from behind, can be swung up and down, and at the same time, its image point α' moves. The light beam wanders up and down on the object and image sides in this process. Only two cross-sections of the beam remain at rest: The aperture B which serves as entrance pupil, and its image B', the exit pupil. To provide a marker, one can cover the upper edge of the entrance pupil with a red glass filter, and its lower edge with a green glass filter. Then the lower edge of the exit pupil appears red, and its upper edge appears green. We thus see how B' arises as the image of B. In contrast, the image point α', which is moving up and down, remains colorless; it contains both red and green parts of the beam, which are complementary colors, so that together, they give white light (for the demonstration, we use a cylindrical lens).

edge of B to the upper edge of B'. They show that B' is the image of B. Often, instead of a transmission aperture, a mirror is used (reflection aperture; Fig. 18.9, top). Example: The mirror is attached to the rotating coil of a sensitive galvanometer, and the lens as objective of a telescope is used for taking scale readings.

The aperture B is imaged actual size (B') in Fig. 18.9 by choosing the distance of the aperture from the lens in the drawing to be $2f$ (compare Eq. (16.13) and Fig. 18.1). When the aperture B is shifted towards the lens, the exit pupil shifts to the right, and at the same time, it increases in size. If the aperture B is placed in the object-side focal plane, then its image becomes virtual and lies at infinity to the right side of the figure; then the common cross-section of the image-side light beams, i.e. the exit pupil, is also at infinity on the right. Thus, the image-side principal-ray angle of inclination $\varphi' = 0$ and the the optical path is *telecentric on the image side*. Examples are shown in Figs. 23.1 and 24.20. The optical path of the light beams sketched in Fig. 18.9 and the positions of the two pupils there can be rather impressively demonstrated in the experiment. Details are given in Fig. 18.10 and its caption.

3. In Fig. 18.11, there is an aperture B *behind* the lens but within the image-side focal length f'. B' is its virtual image. This aperture image B' acts as entrance pupil. Although it is *behind* the image, B' delimits the usable light beams on the object side (opening angle ω). The aperture B itself acts as exit pupil, it delimits the image-side light

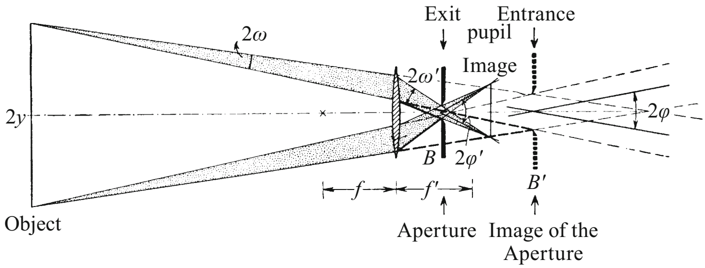


Figure 18.11 As in Fig. 18.9: The exit pupil here is a circular aperture B located on the image side. Its virtual image B', likewise located on the image side, acts as entrance pupil.

beams (opening angle ω'). Once again, some heavy black (dashed) rays indicate that B' is a (virtual, upright) image of B.

4. Frequently, for demonstrations we use an illuminated opening as object; it serves as a light source which is sharply bounded and with a well-defined size and shape (Sect. 16.2). However, it is often difficult to place the lamp close enough to this opening, and the diameter of the lamp or the lens used for imaging is also often too small. In such cases, we can make use of an *illumination lens C*, called a *condenser*, between the lamp and the opening. We demonstrate how a condenser is used with an example in which an illuminated slit serves as a line-shaped light source. In Fig. 18.12, assume that the diameters of both the imaging lens L and of the primary light source (e.g. the crater of the arc lamp) are small. Nevertheless, we want to image the slit along its whole length, so that it appears *uniformly* bright. Then the condenser C has to cast an image of the lamp onto the imaging lens. This image of the lamp may itself be smaller than the area of the lens; in that case, the lens bezel no longer acts as entrance and exit pupil for the image formation, but instead the image of the lamp serves this function. If we make the lamp's image equal to or larger than the area of the imaging lens, we obtain the largest possible opening angle ω' and thus the maximum possible irradiation intensity in the image.

In many cases, as in Fig. 18.12 and some others (e.g. Figs. 18.33 and 24.20), in addition to the *imaging* optical system, there is also a system for the *illumination*. In these cases, pupils are indicated for the *whole* optical arrangement. Thus, for example, in Fig. 18.12, the area of the lamp (crater of the electric arc) is the entrance pupil, and its image on the lens L which forms the main image is the exit pupil. When several apertures are present, it is important to distinguish the one which limits the pupils from all the others. It is called the *aperture stop* (Sect. 16.1).

The facts which are elucidated in Figs. 18.9–18.12 may be summarized as follows: The light beams which actually propagate (going from an object point to the lens and from the lens to the corresponding image point) are determined by the entrance and exit pupils. These

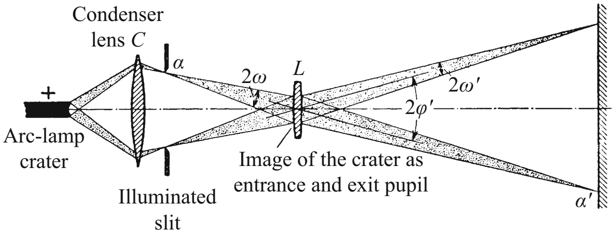


Figure 18.12 Delimiting of the light beam and the positions of the pupils when imaging with a condenser lens which uniformly illuminates an opening (here, a wide slit); for the positions of the pupils, see also the above small-print paragraph)

Part II

pupils are either physical apertures (e.g. openings, lens bezels, mirrors), or they are the light-emitting surfaces of a light source, or else they are images of an aperture or of the light source. These images can be real or virtual. The *entrance pupil* is the common cross-section of all the light beams coming from the object side, while the *exit pupil* is the common cross-section of all the light beams passing through the image side. The diameters of the pupils determine the usable opening angles ω and ω'. The centers of the pupils practically always lie on the symmetry axis of the lenses (the optical axis). Then these center points are the intersections of the object-side and the image-side *principal rays* and are thus at the apex of the principal-ray angles of inclination φ and φ'.

In using lenses, one therefore needs to keep two things clearly separated: The rays used in the geometrical construction and drawn upon the patient sheet of paper (e.g. Fig. 18.1), and the real, usable light beams, which are delimited by pupils. Of course, we could also construct the images drawn in Figs. 18.9 ff. using the scheme shown in Fig. 18.1. The reader may even wish to check the validity of Fig. 18.9 or 18.11 in this manner. However, we must never confound the rays drawn with the axes or the boundaries of the real light beams which in fact are observed in the experiments.

"In using lenses, one therefore needs to keep two things clearly separated: The rays used in the geometrical construction and drawn upon the patient sheet of paper, and the real, usable light beams, which are delimited by pupils".

A precise insight into the delimiting of light beams by pupils is absolutely indispensable for all optical apparatus and experimental setups. The boundaries of the light beams play a decisive role in the construction of optical systems, e.g. the compound lenses with which we reduce the unavoidable imaging aberrations to an acceptable level.

The bounding of the light beams determines the radiant power which can be transmitted by an optical system, and therefore (using the language of everyday life), all questions of brightness.

In all optical instruments, e.g. microscopes and telescopes, the light-beam boundaries limit the field of view, the depth of focus, the perspective and the usable magnification.

Finally, the light-beam boundaries also determine the *resolving power* of all optical instruments, not only that of the eye and of telescopes (Chap. 17). They limit e.g. in the microscope the smallest length within the object which can still be distinguished; in a spectral apparatus the smallest differences in wavelengths which can still be separated, etc.

18.4 Spherical Aberration

The derivation of the lens formulas in Sect. 16.7 presumed that the light beams were thin and near to the optical axis. The approximation used was that the ratios of the sines of two angles could be replaced by the ratio of the angles themselves (small-angle approximation). For larger angles, the ratio of the angles is larger than the ratio of their sines. Thus, for example, $\sin 90° = 1$, $\sin 45° = 0.7$, so that while $90°/45° = 2$, in contrast, $1/0.7 = 1.4$. This fact is the reason for the occurrence of imaging aberrations, even when monochromatic light is used, as soon as the ideal situation of very thin light beams near to the optical axis (i.e. paraxial rays) no longer applies.

In Sects. 18.4–18.9, we deal with the most important imaging aberrations, giving a brief summary in each case. Up to Sect. 18.8, we assume that red-filter light is being used. Furthermore, when the contrary is not expressly stated, we assume that the light beams are bounded by a circular aperture, in the simplest case by the lens bezel. The center point of this aperture is presumed to lie on the symmetry axis of the lens, i.e. the optical axis of the setup.

We start with spherical aberrations. The object-side and image-side opening angles ω and ω' were defined in Fig. 16.21. If at least one of them is large, then each single zone of the lens produces its own image point P' from a single object point P that lies on the lens axis. These image points no longer overlap, but instead form a series of image points along the optical axis. (The individual zones of the lens thus have slightly differing focal lengths.) This imaging error is called “spherical aberration”.

In order to demonstrate spherical aberration, in Fig. 18.13 (left-hand part), we place the object point (the crater of an arc lamp) far to the left on the optical axis. In the plane of the page is a matte-white screen which is grazed by the light beams. In addition, we place a diaphragm with four openings near the lens; it allows four nearly parallel light beams to pass. Their intersection with the plane of the page shows an outer and an inner pair of beams. The inner pair passes through the neighborhood of the center of the lens, while the outer pair passes through a lens zone near its outer rim (lens bezel). The intersection point of the outer beam pair occurs *before* the intersection point of the more central pair, as seen in the direction of the light propagation: This lens is “spherically undercorrected”.

Fig. 18.13 (right-hand part) shows the corresponding experiment using a concave lens. The pair of beams from near the rim of the lens

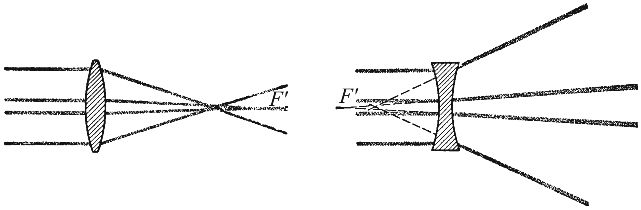


Figure 18.13 The demonstration of spherical aberration using cylindrical lenses (with their cylinder axes perpendicular to the plane of the page); at left: spherically undercorrected, at right: spherically overcorrected **(Video 18.1)**

intersects, again in the direction of light propagation, *behind* the pair which passes near the center of the lens. This lens is "spherically overcorrected".

In order to correct for spherical aberration, we see that a suitable combination of convex and concave lenses can be used. Spherical aberration can always be corrected only for certain values of the object and image distances. For telescope and camera objectives, one chooses an object at infinity. Microscope objectives are corrected for an object located close to the object-side focal point.

For many imaging applications (e.g. in the lecture room), the object and image distances are very different. In such cases, it often suffices to use a simple plano-convex lens: We let the light beam with the greater opening angle be on the *planar* side of the lens. Then the rays pass through the outer zones of the lens with approximately "minimum deflection" (Sect. 16.6) (since the object distance is much less than the image distance, the path of the rays through the rim zone of the lens is more symmetric when the planar side of the lens is oriented towards the object). This simple trick strongly reduces the spherical aberration (compare Fig. 18.15, figure caption).

Not only convex lens surfaces, but also planar surfaces can give rise to spherical aberration. As a result, microscope objectives can be corrected only for a prescribed cover-glass thickness when light beams with a wide opening angle are used (i.e. beams with a large aperture; cf. Sect. 18.12). This must be taken into consideration when using a particular objective.

Video 18.1:
"Spherical aberration"
http://tiny.cc/1eggoy
In the video, the demonstration of spherical aberration is carried out using a small glowing filament which is imaged onto the wall of the lecture room. Various metal panels placed in front of the lens permit us to use either only the central portion of the lens, or only the zone near its outer rim. When the light passes only through its central portion, the image is sharply focussed. When only the rim zone is used, the image is completely unrecognizable. Only when the experimenter moves the image plane closer to the lens by using a white cardboard screen can the image of the filament be discerned, but it is still of a rather poor quality.

18.5 Astigmatism and Arching of the Image Plane

Spherical aberration occurs even for object points on the optical axis. In general, however, a given object point P will be at some distance from the optical axis, for example when one wishes to photograph a landscape. In this case, simple lenses can produce an image *point* only with extended (planar) light beams or "light fascicles". Such beams are obtained by using a very narrow slit in a diaphragm. The

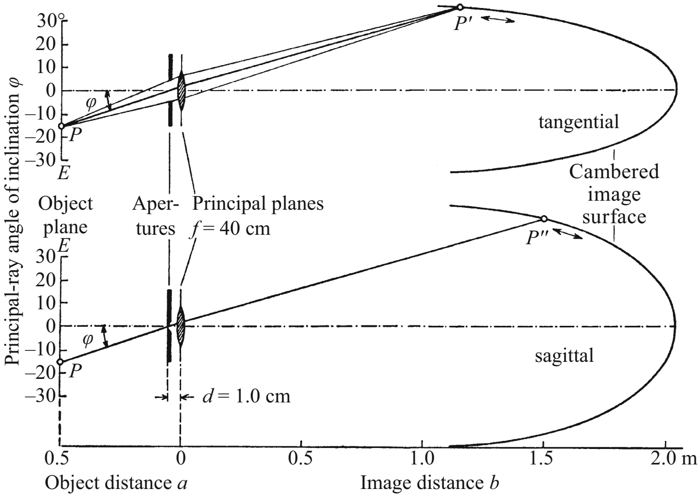


Figure 18.14 The demonstration of astigmatism and image-surface cambering (also known as "arching"), using very thin, planar light beams defined by a slit (the experimental setup is shown as a plan view (from above), not as a side view). In the upper part of the figure, the planar light beam is in the plane of the page. Below, the planar light beam is perpendicular to the plane of the page; we see only its principal ray. The lateral rays which emerge from the slit pass below and above the plane of the page. If we replace the slit by a round opening, then we will no longer obtain image points, but rather "image streaks". In order to make them visible to a larger audience, we use a lens of large diameter, around 10 cm (**Video 18.2**).

Video 18.2: "Astigmatism"
http://tiny.cc/reggoy
In order to give an impression of the distortions of the image points caused by astigmatism, the crater of an arc lamp is projected in the video onto the wall of the lecture room using a lens with a long focal length and large diameter. We see a small, relatively sharp disk of light. If the lens is rotated around a vertical axis, then only light beams which pass obliquely through the lens can contribute to image formation. The light disk in the image plane then takes on long, stretched-out shapes (0:25 min). Finally, with the lens slanted, the lamp is moved closer to it. During its motion, both horizontal and vertical image streaks appear in turn (0:55 min). When, however, the distance to the lamp is varied while the lens is oriented normally ($\varphi = 0$) (1:28 min), only the *size* of the circular "image point" changes.

long axis of the slit can lie either *in* the incident plane (tangential) or *perpendicular* to the incident plane (sagittal). The former case is shown in the upper part of Fig. 18.14, while the latter is shown below, as plan views.

For the demonstration, we change the angle of inclination φ of the principal rays arriving from the object side, by sliding the object point P (for example the crater of an arc lamp) along a rail E in the object plane. At the same time, we determine the distance b of the screen at which a sharply-focussed image point appears. (The necessary shifts in distance, often several meters, are most conveniently managed by using a cart, similar to that shown in Fig. 18.15b.)

For every angle of incidence φ, both orientations of the slit each give *one* rather sharp image point, P' and P''. They lie at different distances from the lens. Only in the limiting case of $\varphi = 0$ are both of them found at the same distance. The difference of the distances of these two image points is called the *astigmatism*. The set of all the image points P' and P'' for the planar light beams lying in the plane of incidence and perpendicular to it each form a rotationally-symmetric concave surface. The two image surfaces are cambered and touch each other for the limiting case of $\varphi = 0$, that is when the object and the image point are on the optical axis.

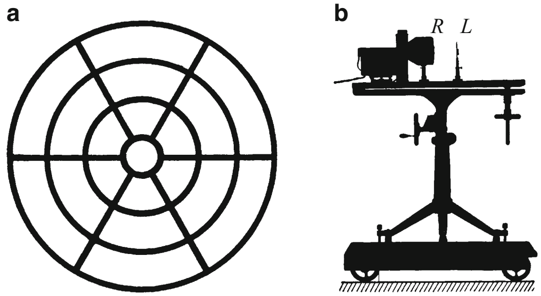


Figure 18.15 A wheel with several rims drawn onto a matte glass plate (the spokes and rims are transparent, while the rest of the area is opaque) is an excellent object for testing a lens for astigmatism and image-surface cambering (about 1/2 actual size). Among others, a demonstration experiment using a plano-convex lens of around 13 cm focal length and 4 cm diameter is very impressive. If the planar surface is oriented towards the object, there is a strong image-surface cambering and astigmatism. The spokes and the rims are sharply focussed at different distances from the lens. It is expedient to put the optical bench onto a cart, since it must often be moved over several meters in order to sharply focus either the spokes or the rims, using beams either from near the center of the lens or from near its outer edge. If the convex surface of the lens is oriented towards the object, the image surface is surprisingly flat, but now the strong spherical aberration makes the rims appear one-sidedly washed out towards the center of the image. (R = wheel figure (object), L = lens.)

If we replace the slit by a circular aperture, for example the lens bezel, then we observe an additional complication. At the locations of the two image points P' and P'', we see at the same time two nearly linear formations which are perpendicular to each other: Each of the two image points has degenerated into an *image streak*. In P', the streak is perpendicular to the plane of incidence, while at P'', it lies in the plane of incidence.

To explain these image streaks, we can refer to Fig. 16.17. In each direction, a lens which is struck by light beams at a *slanted* angle can be replaced by two cylindrical lenses with different curvatures for light beams with a small opening angle. Therefore, with slanted incidence, the same effects must be observed as in Fig. 16.17b with incident beams parallel to the optical axis.

When lenses in meniscus form (concavo-convex, e.g. Fig. 18.7 a) are used, at a suitable position of the aperture, the order of the concave-cambered image surfaces is exchanged. That means that the image surface nearer to the lens is due to the extended light beam which is perpendicular to the plane of incidence (that is, the sagittal beam). By combining convex and meniscus lenses, one can cause the two concave surfaces to very nearly merge, thus strongly reducing the astigmatism. In addition, the common concave surface can be more or less flattened out, so that the camber of the image surface is reduced to an acceptable level. Such compound lenses (objectives) with strongly reduced astigmatism and a roughly planar image surface are called *anastigmatic*.

To test an objective for image-surface cambering and for the degree of its astigmatism, the degeneration of the image points to image streaks is employed: One places a drawing of a wheel, with spokes and a rim, perpendicular and symmetric to the optical axis. When the correction is imperfect, either only the spokes or only the rim can be brought into sharp focus. Often, the drawing shows several concentric rims (Fig. 18.15). With well-corrected objectives, even the outer rims as well as the spokes should be projected in sharp focus onto a flat screen.

18.6 Coma and the Sine Condition

In *spherical aberration*, the object point lies on the optical axis, i.e. the angle of inclination φ of the principal ray is zero. The cross-sectional area of the beam retains its *circular symmetry* on the image side. Correspondingly, the image point degenerates to a *circular disk* when the opening angle becomes too large.

However, with *astigmatism*, the object point lies outside the optical axis, i.e. the principal ray (which passes through the center of the aperture) has a finite angle of inclination ($\varphi > 0$). It is no longer perpendicular to the surface of the lens. As a result, the cross-sectional area of the beam acquires an *elliptical* shape on the image side, so that the two image points become image streaks even at small opening angles ω.

Now, however, if at least one of the opening angles is large, then the cross-section of the beam retains only its *symmetry with respect to the plane of incidence* on the image side. The image points acquire a "tail" which effectively broadens them and reduces their intensities; they degenerate to a *coma*. A coma can occur even when the spherical aberration has been corrected. Even then, at large opening angles ω, a small angle of inclination φ of the principal ray changes the focal lengths of the individual zones of the lens. Therefore, each lens zone projects the image in a different size onto a surface element that is perpendicular to the optical axis. This however would make the use of light beams with a large opening angle impossible for microscopes, telescopes and other optical instruments. It is thus not sufficient, as shown above in Sect. 18.4, to reduce the spherical aberration for only one object and one image point on the optical axis. Rather, this must be done for other object and image points which are *not* on the optical axis. This can be accomplished by following a certain rule for the ratio of the opening angle ω on the object side to the opening angle ω' on the image side. These angles must fulfill the *sine condition*:

$$\frac{n \sin\omega}{n' \sin\omega'} = \frac{\Delta y'}{\Delta y} = \text{const} \quad \text{or} \quad \frac{\sin\omega}{\sin\omega'} = \frac{\Delta y'}{\Delta y} = \text{const}\,; \qquad (18.2)$$

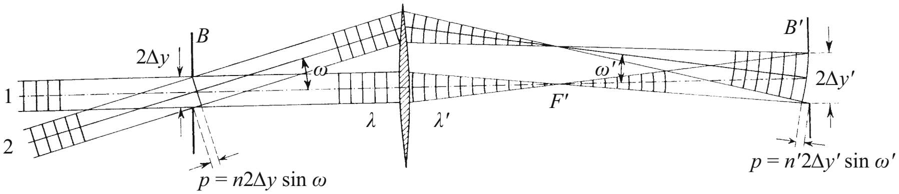


Figure 18.16 The derivation of ABBE's sine condition. An aperture B which serves as entrance pupil is imaged as the exit pupil at B'. For the illumination of B, we suppose that there is an extended light source some distance to the left. The indices of refraction n and n' on the object and the image sides are taken to be equal in this sketch; therefore, the wavelengths λ and λ' are drawn to be the same.

the latter expression holds when the indices of refraction n and n' on both sides of the lens are equal.

To derive the sine condition most simply, we refer to Fig. 18.10. There, the entrance pupil was formed by an aperture B, and its image served as exit pupil B'. The same can be seen in Fig. 18.16; however, here, the light source is not an illuminated opening, as in Fig. 18.10, but instead a light source with a large surface area, far to the left in Fig. 18.16. We have drawn two collimated beams which each originate at a point on the distant light source. Some wavefronts are indicated in each beam. One of the beams passes through the center of the lens, while the other passes through a zone near the rim of the lens. The central axes of these two beams make an opening angle ω on the object side, and an angle ω' on the image side. Both beams intersect the image plane with the same cross-section (the diameter of the image is $2\Delta y'$). The wavefronts are perpendicular to the rays which form the edges of the beams; they appear as straight lines on the object side. On the image side, we can see that their curvature just in front of the image plane is small enough to be neglected. The last wavefront which is drawn in on the right can be treated as flat. Then from Fig. 18.16, we can read off the optical paths $p = n2\Delta y \cdot \sin\omega$ and $p' = n'2\Delta y' \cdot \sin\omega'$. Since for image formation, the optical paths of all the rays that belong to corresponding object and image points must be equal[C18.3], we must have $p = p'$. This leads to the sine condition (18.2).

An optical arrangement for forming an image which obeys the sine condition is called *aplanatic*. Such a lens can thus image a certain *surface element* which is perpendicular to the optical axis (and not just an *object point*) with wide-angle beams. However, a lens can deliver an aplanatic image only for certain object and image distances, which must be predetermined during its fabrication.

C18.3. This follows from FERMAT's principle; see Comment C16.6. A detailed derivation can be found for example in P. Drude, "*Lehrbuch der Optik*", Verlag S. Hirzel 1900, Chap. III.9. This book, in its Chap. III.10, also contains a derivation of the tangent condition which is mentioned in the following section.

18.7 Geometric Distortion

Spherical aberration, astigmatism and coma impair the quality of the image points. They are *aberrations in image definition*. In addition, there are *positional aberrations*. They give rise to a camber of the

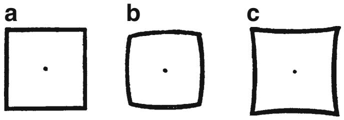


Figure 18.17 A square a centered on the optical axis has a barrel-shaped distortion at b and a pillow-shaped distortion at c (a is drawn with about 10-fold magnification on a matte glass plate, preferably as a light figure on a dark background)

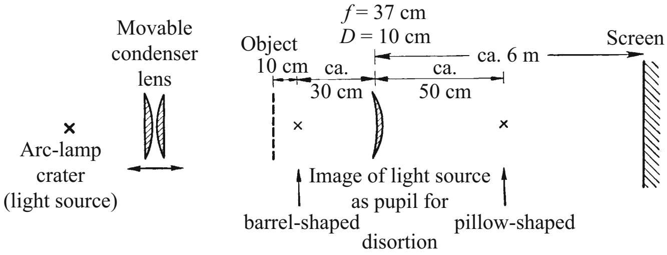


Figure 18.18 The dependence of the distortion on the position of the pupil (the object is for example a square lattice). D is the lens diameter

image surface and distortion of the image: A square is distorted either to a pillow shape (Fig. 18.17c) or to a barrel shape (Fig. 18.17b). An image is free of distortion at a fixed position of the pupil when the principal-ray angles of inclination φ and φ', e.g. in Fig. 18.11, fulfill the condition $\tan\varphi'/\tan\varphi = \text{const}$ (*tangent condition*; see Comment C18.3). Often, however, the position of the pupil is not fixed; it varies for different zones of the objective. Then the distortion can no longer be corrected. An example of this situation can be seen in Fig. 18.18.

18.8 Chromatic Aberration

The focal length of a lens depends not only on the shape of the lens surfaces, but also on the index of refraction n of the lens material. The focal length f is proportional to the reciprocal of $(n - 1)$ (compare Eq. (16.12)). All materials which can be used to fabricate lenses, i.e. glasses, crystals and plastics, exhibit dispersion; in general, their index of refraction increases within the visible spectral region as the wavelength decreases (cf. e.g. Table 16.2). Therefore, the lens has a different focal length for each wavelength.

The focal length determines both the *position* of the image and also its *size*. Therefore, there is a chromatic aberration both in the image *position* as well as in its *reproduction scale*. (In addition, the other aberrations acquire a dependence on the wavelength which is of practical significance.)

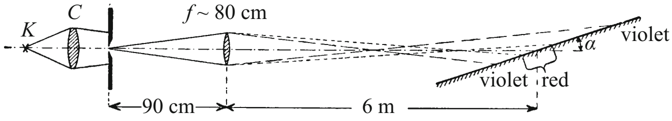


Figure 18.19 A demonstration of the chromatic aberrations of the image *position* and its *reproduction scale* for image formation by *thin* lenses. Only with a thin lens is the position of the principal planes practically independent of the wavelength. Therefore, only for a thin lens does the same *position* of the focal point imply equality of the focal lengths and of the *reproduction scale* which they determine (the angle of inclination α of the screen is about 10°) **(Video 18.3)**.

Video 18.3:
"Chromatic aberration"
http://tiny.cc/seggoy
The video shows a special example: An arrow formed from holes illuminated from behind is projected onto the wall of the lecture room using a glass lens with a strong dispersion. The holes are illuminated by an arc lamp with a condenser lens and an auxiliary lens, giving a weakly convergent beam, so that the slender light beams from the arrowhead and from its tail strike only the rim zone of the lens, while those from the middle of the arrow strike only the center zone of the lens. The image, which is upside down, shows colored borders for the arrowhead and tail. The dispersion between the refraction of red and blue light, which gives rise to chromatic aberrations, increases near the rim of the lens. Red light (with a longer wavelength) is less strongly refracted than blue light (with a shorter wavelength). Slipping a red filter into the optical path allows us to compare monochromatic and polychromatic image formation. We can see that the position of the red borders fits onto those of the full red light disks.

Both of these chromatic aberrations can be readily demonstrated using a simple eyeglass lens. We use the lens to image a slit onto a distant screen which can be slid along the direction of the light beam, and put alternately a red or a blue filter in front of the slit. In order to focus the blue image, the screen must be placed considerably closer to the lens than for the red image: "chromatic aberration of the image position". The blue image is also about seven-eighths the size of the red image: "chromatic aberration of the image size" (i.e. of the reproduction scale). If the screen is tilted to the side (see Fig. 18.19), we see a broad, colored band instead of the image of the slit. The non-physicist would be inclined to call this band a 'spectrum', just like a rainbow. The physicist can admit of only a remote similarity in both cases.

As with all imaging errors, chromatic aberrations can be reduced but not eliminated. For this reduction, called "achromatization", one uses in practice at least two lenses. For the achromatization of the image *position*, convex and concave lenses made of *different materials* are required. To achromatize the *reproduction scale*, two convex lenses made of the *same* type of glass can be used with parallel light beams. Their axes must coincide and their spacing should be equal to half the sum of their focal lengths. *Achromatic lenses consisting of two lenses of the same type of glass* can be found for example in the ocular lenses of binoculars (Fig. 18.20): They allow parallel light beams of different wavelengths to pass into the eye between the two focal points F' with a *parallel displacement* relative to one another. A parallel displacement (just as in the mirror prisms in Sect. 16.11) leaves the image on the focal plane of the eye unchanged.

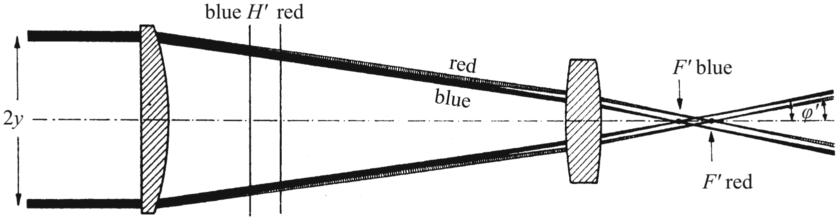


Figure 18.20 A demonstration of an *achromatic* ocular *using two lenses made of the same type of glass*. The second lens makes the initially divergent rays of different wavelengths parallel. This allows them to be focussed together on the focal plane of the eye.

18.9 Achievements of Optical Technology. The SCHMIDT Mirror

Optical technology has made it possible to reduce lens aberrations, sometimes individually, sometimes in combination. The main tool for this purpose is the use of compound lenses. These consist of a series of individual lenses with spherical surfaces and a common optical axis. Non-spherical surfaces are relatively seldom used, for example parabolic mirrors for telescopes and searchlights, or non-spherical lenses as condensers for projection apparatus.

Every lens and every mirror must be precisely adapted to its special use. The objective lens of a microscope must meet quite different requirements from those of a telescope objective. A magnifying lens used for reading a scale has a different construction from that of a magnifying glass for examining photographs, and so forth. General methods for reducing the individual lens aberrations have long been known, but the optimal solution of each special problem necessitates in general a numerical calculation which makes skillful use of the various types of glass available. Technology has made admirable progress in this field and has thereby provided considerable support for research. A particular success was for example the development of the coma-free SCHMIDT mirror (BERNHARD VOLDEMAR SCHMIDT, 1879–1935). The principle of this important invention is explained in Fig. 18.21 and in Comment C18.4.

The development of objectives with continuously-variable focal lengths ("zoom lenses") is also remarkable. They consist of at least two individual lenses, which can be displaced by differing amounts relative to the image plane. Jokingly called *rubber lenses*, they can change their focal lengths and reproduction scales continuously by up to a factor of 5 (for example *during* the filming of a motion picture), with a focal ratio of nearly 1:1 (see Sect. 18.15).

C18.4. B.V. SCHMIDT, who lost his right hand at age 11 in an accident while experimenting with gunpowder, invented the telescope which bears his name at the Hamburg Observatory in 1931. He developed the correction plate (lens) and at the same time a process for grinding the lens. He used the lens as the lid of an evacuated vessel (the vacuum is on the right side in Fig. 18.21, Part b), in order to grind it to a concave shape while it was elastically distorted by air pressure, thus producing the "peculiar" profile shown in Fig. 18.21. His invention made it possible for the first time for astronomers to obtain wide-angle celestial photographs.

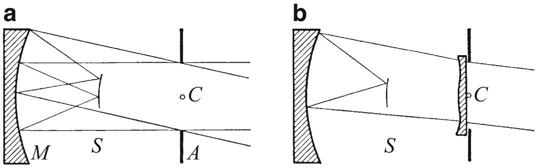


Figure 18.21 The principle of the coma-free SCHMIDT mirror. Part a: At the center of curvature C of a concave spherical mirror M is the aperture A which serves as entrance and exit pupil. The image points lie on the spherical surface S whose center of curvature is at C. They have no chromatic aberrations, but when the diameter of the aperture A is large, there is an undercorrection of the spherical aberration of the mirror. Part b: SCHMIDT compensated this by a spherical overcorrection using a glass plate[C18.4], whose special profile is sketched here (greatly exaggerated). After the spherical aberration had been corrected in this manner, opening angle ratios of 1:1 were readily attained.

18.10 Increasing the Angle of Vision with a Magnifying Glass or a Telescope

In considering the eye, we may continue to employ the analogy with a photographic camera for the time being. The eye can accommodate, i.e. it can produce sharply-focussed images of objects at various distances. In a camera, for this purpose the distance between the rigid glass lens and the image plane (detector array, film or photographic plate) is varied. The eye, in contrast, changes the curvature of its lens surfaces by muscular contraction, and thereby changes the focal length f' of its elastically-deformable lens.

The range of accommodation for a normal eye stretches from an arbitrarily great distance ("infinity") down to the "near-point distance". This closest distance for sharp focus, requiring the strongest accommodation, can be below 10 cm in children. For adults between age 20 and 40, one finds near-point distances of around 20–25 cm, and so forth. – However, very strong accommodation is uncomfortable. For writing, reading and fine handwork, in general a distance of around 25 cm is preferred. This preferred working distance is called (rather unfortunately) the "distance of most distinct vision".

An important but complicated process is *three-dimensional* vision, whether directly, in a mirror, or through a water surface. Its essential aspects are dealt with in physiology.

In describing even elementary optical observations, one fact must be kept firmly in mind: A single object point, *alone* in the field of view, can be physically localized only by *two* eyes.[C18.5] *One* eye can always determine only the *direction* in an unknown environment in which we see such an object point L, but never its distance (compare Fig. 18.22).

C18.5. See also Sect. 1.4 in Vol. 1: The stereoscope.

With the aid of Fig. 18.23, we define the *angle of vision*. The angle of vision must not fall below a certain minimum value (about 1 minute of arc) for well-known reasons (Sect. 17.3); otherwise, the eye can no longer resolve the point or separate it from its surroundings.

How can we increase the angle of vision, how can we make previously indiscernible details in an object visible? Answer: We approach the object more closely. How closely can we approach it? Normally down to about 25 cm, the *distance of distinct vision*, with comfort. At still closer distances, persons with normal vision find accommodation difficult, and without accommodation, they would see only an unfocussed image. However, the curvature of the eyeball can be assisted by placing a convex lens in front of it (Fig. 18.24). Then the observer can approach the object more closely, e.g. to 12 cm, without any effort of accommodation, and still see a sharp image. This closer approach roughly doubles the angle of vision in comparison to the distance of distinct vision ($\varphi_{\mathrm{wi}}/\varphi_{\mathrm{w/o}} \approx 2$). Or, in other

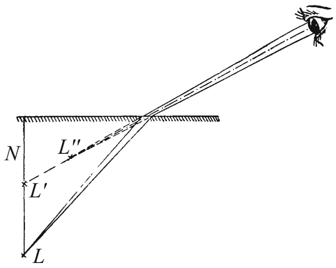


Figure 18.22 An object point L under water is viewed by two eyes and appears to be *raised vertically* to the point L'. *This cannot be understood in terms of this construction which is carried out with only one eye*. That would lead us to an object point L'' which appears to be shifted towards the observer. Instead, we have to carry out the construction separately for both eyes. Then the two planes of incidence intersect in the vertical (normal) line N, and L' also lies on this normal as the point of intersection of the two axes of the light beams.

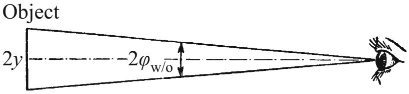


Figure 18.23 The definition of the angle of vision 2φ without any instruments or visual aids, denoted by $2\varphi_{w/o}$

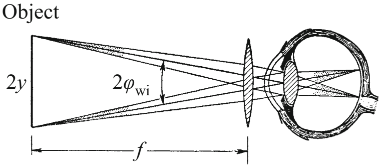


Figure 18.24 Increasing the angle of vision (to $2\varphi_{wi}$) by using a magnifying glass or loupe. Instead of using the image formed by the lens in the iris opening of the eye, the opening itself serves to good approximation as the entrance and exit pupil. The differences between the principal and nodal points are neglected here. A more detailed treatment of the action of a magnifying glass would go beyond the framework of this book.

words, we have placed a *magnifying glass* (or *loupe*) in front of the eye. A magnifying lens with a still stronger curvature would permit approaching the object to 5 cm, then its magnification is about 5x, and so forth. The purpose of a magnifying glass is thus to increase the angle of vision by permitting a closer approach of the eye to the object. The magnification of the lens is here not a constant in the physical sense. It increases with the age of its user owing to the gradually decreasing ability of the lens of the eye to accommodate.

Experienced observers always use a magnifying glass with their eye relaxed, that is accommodated to ‘infinity’. They place the object in the focal plane of the magnifying glass (Fig. 18.24). Then the

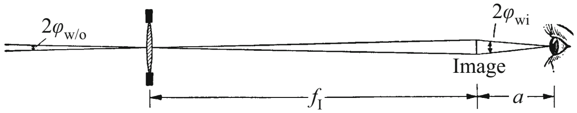


Figure 18.25 Increasing the angle of vision by using a single-lens telescope, here as a simple sketch showing only the principal rays. One can imagine a frosted glass screen in the image plane; it is however not necessary. (Numerical example: $f_{\text{I}} = 4\,\text{m}$, distance of the eye from the image $a \approx 20\,\text{cm}$, magnification $f_{\text{I}}/a = 20\text{x}$)

light beams coming from the individual object points enter the eye as parallel beams. The lens of the eye refracts the beams so that they are convergent and focusses their narrowest points of convergence as image points onto the retina of the eye.

Often, it is not possible to approach the object more closely (an airplane in the air, the moon, etc.). Then one can use an objective lens to project an image of the object. The image is of course much smaller than the object itself, but the observer can approach it to within about 25 cm (the distance of distinct vision) and thus increase the angle of vision. This is the principle of the single-lens telescope shown in Fig. 18.25. By adding a magnifying glass in front of the eye of the observer, the approach can be made much closer and the angle of vision still larger. This is a two-lens telescope as shown in Fig. 18.26. The objective and the magnifying glass ("ocular") are connected by a tube (which holds them fixed and shields unwanted stray light from the surroundings). The telescope thus also simply serves to increase the angle of vision. The magnification or "power" x of a telescope is defined as the ratio "angle of vision with" over "angle of vision without" the instrument (its measurement procedure will be described in Sect. 18.13).

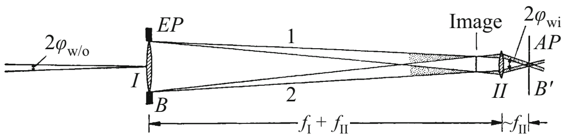


Figure 18.26 Adding a magnifying glass as ocular (lens II, focal length f_{II}) allows the eye of the observer to be brought closer to the image and thereby increases the angle of vision still further. On the object side, we again show only the principal rays which come from the rim of the object; on the image side, however, the corresponding light beams are drawn in. The objective lens bezel serves as entrance pupil EP, and its real image B', projected by the ocular, is the exit pupil AP. We can imagine this image to just fill the opening of the iris in the observer's eye, as in Fig. 18.24. The eye should be relaxed when using a magnifying glass or ocular, that is the entering light beams should be parallel. (The rays 1 and 2 show that B' is an image of B.) The magnification is roughly equal to the ratio of the focal lengths: $x \simeq f_{\text{I}}/f_{\text{II}}$ (see Sect. 18.13).

The type of telescope sketched in Fig. 18.26 was suggested in 1611 by Johannes Kepler (1571–1630), and is called an "astronomical" telescope. it shows an upside-down image of the object. There are various methods for obtaining an upright image, for example additional lenses or mirror prisms (Fig. 16.35) between the objective and the ocular.

18.11 Increasing the Angle of Vision with a Projector or a Microscope

The generally well-known optical instruments 'projector' and 'microscope' serve – like the telescope and the magnifying glass – to increase the angle of vision. The two instruments are essentially similar in principle. In both, the object is placed just outside the object-side focal point of an objective lens. This lens then projects a strongly enlarged image of the object. The image can be observed on a screen (real image).

When it is sufficiently large, the image can be clearly observed with a sufficiently large angle of vision even by observers located some distance away; this is the function of a projector (slides, motion pictures!).

The microscope, in contrast, is intended for individual observations of small objects. The image projected by the objective lens is at the upper end of the microscope tube ("tubus"). The observer uses the ocular (a magnifying glass) to approach the image closely and thus observes it under a large angle of vision. The magnification or "power" is defined for the microscope again as the ratio "angle of vision with" over "angle of vision without" the instrument.

In order to measure the magnification of a microscope, we lay a millimeter ruler on the object table of the microscope and allow a piece of it to protrude over one side, e.g. to the right. Then we look through the microscope with the left eye, while the right eye looks at the ruler directly. We can readily bring the two fields of vision together. We see for example 1 mm through the microscope over 130 mm on the directly-observed ruler. Then the magnification is 130x.

18.12 Resolving Power of the Microscope. The Numerical Aperture

The discussion in Sect. 17.3 on the resolving power of lenses holds equally well for the microscope as for the eye and the telescope. The angular spacing of two still separately visible object points – called

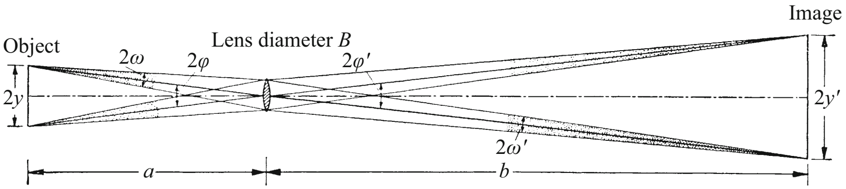


Figure 18.27 The resolving power of the microscope. Here, the same conditions apply as in Sect. 17.3 for the derivation of Eq. (17.1) for the eye and the telescope. On the right side of the lens, the light beams have in reality nearly parallel boundaries: The object points on the left side are nearly in the focal plane of the objective. In the drawing, for clarity the object distance a is shown overly large and the image distance b overly small. Furthermore, in the sketch, the indices of refraction n and n' on the object side and the image side are shown as being equal. The drawing thus applies to a microscope without any immersion fluid.

2φ in Fig. 18.27 – must not be smaller than the angle α calculated from the equation

$$\sin\alpha = \frac{\lambda}{B}\,. \tag{16.23}$$

Then

$$\sin 2\varphi_{\min} = \frac{\lambda}{B} \quad \text{or, from Fig. 18.27,} \quad \frac{2y'}{b} = \frac{\lambda}{B}\,. \tag{18.3}$$

However, in the case of the microscope, the smallest resolvable angle $2\varphi_{\min}$ is less interesting than the smallest separable spacing of two object points; that is, in Fig. 18.27 the distance $2y_{\min}$, measured in units of length.

For its calculation, we read off from Fig. 18.27 for the image-side (small) opening angle ω' the following relation:

$$\sin\omega' \approx \omega' \approx \frac{B}{2b}\,. \tag{18.4}$$

Furthermore, in the microscope, the *sine condition* must be fulfilled, i.e. the image-side opening angle ω' must be related to the object-side opening angle ω by the equation

$$\frac{n\sin\omega}{n'\sin\omega'} = \frac{2y'}{2y}\,. \tag{18.2}$$

Equations (18.2), (18.3) and (18.4) can be combined to yield

$$2y_{\min} = \frac{\lambda}{2n\sin\omega}\,, \tag{18.5}$$

in the case that the space between the object and the microscope objective is filled with an "immersion fluid" (water or oil) with the index of refraction n, and $n' = 1$.

This shows that the resolving power of a microscope is determined by two quantities: First, by the wavelength λ of the light, and second by a quantity which characterizes the objective, ($NA = n \sin\omega$), called the "numerical aperture" . In it, ω is the opening angle of the light beams accepted by the objective and n is the index of refraction of the material (air or immersion fluid) between the objective and the object (for example stained thin sections of biological material).

Optical technology has been able to attain values of the numerical aperture $NA = n \sin\omega$ using immersion fluids of up to about 1.4 ($\omega = 70°$, $\sin\omega = 0.94$, $n = 1.5$). The average wavelength λ of visible light is around 600 nm. With these values, from Eq. (18.5) we find

$$2y_{\min} = \frac{600\,\text{nm}}{2 \cdot 1.4} \approx 210\,\text{nm} = 0.21\,\mu\text{m}\,.$$

The smallest spacing of two object points which can still be resolved by high-quality microscopes is thus only slightly less than half the wavelength of the light used.[C18.6] The order of magnitude corresponds to our experience in mechanics. There (Vol. 1, Figs. 12.12 ff.), we produced shadow images of partially-immersed objects using water surface waves. For that simplest type of image formation, we found that the objects must be no smaller than roughly the wavelength of the water waves (cf. Vol. 1, Fig. 12.20).

C18.6. There are more recent developments which permit the size of the resolving power to be reduced below the diffraction limit. These include the optical scanning near-field microscope and the fluorescence microscope (e.g. the STED microscope developed by S. HELL), with which resolving powers (smallest resolvable spacing) of typically less than 30 nm can be attained.

Later, we compare Eq. (18.5) to the coherence condition in Fig. 20.4, in which $n = 1$ is assumed. This will give the equation an intuitively clear interpretation!

Microscopic image formation at high resolution requires that the object-side light beams have a large opening angle ω. This can be seen from the denominator in Eq. (18.5). For *light-emitting* objects (primary light sources) such as a glowing filament, the usable opening angle is limited only by the structural characteristics of the object. In the case of *illuminated* objects (secondary light sources), in contrast, for example the usual stained thin sections observed for biomedical applications, the opening angle is determined by the type of illumination. It is provided by illumination lenses, called "condenser lenses".[C18.7] Figure 18.28 shows two types of construction. At the left, the light passes through the object (a thin section) and into the objective, and thence to the eye of the observer. The object is viewed against a bright background, or with *bright-field illumination*. At the right, in contrast, the light for illumination of the object is kept out of the objective (by total reflection at the surface of the cover glass). Only the light scattered or refracted by the thin section (three small arrows!) can enter the objective. The object then appears bright against a dark background, or in *dark-field illumination*.

C18.7. The function of a condenser lens was already explained in Fig. 18.12.

We are already familiar with bright-field and dark-field illumination from everyday life. We use for example lace curtains with large openings in front of a bright window; this provides *bright-field illumination*. A fine Brussels lace, on the other hand, mounted on

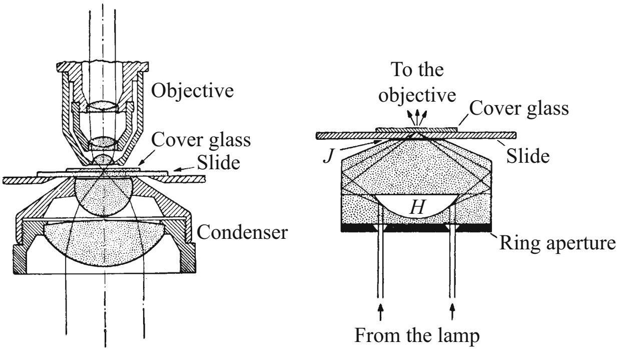


Figure 18.28 Two condenser lens systems. Both provide an extended light source in their focal planes, and this coincides with the plane of the object. The extended light source is obtained by using a convergent lens which images a (pointlike) light source in its entrance pupil. Left: A bright-field condenser. Right: A dark-field condenser with a twofold reflection at the inside surface of the cover glass. H is a cavity, J the immersion fluid (water or oil) for avoiding total reflection at the upper surface of the condenser.

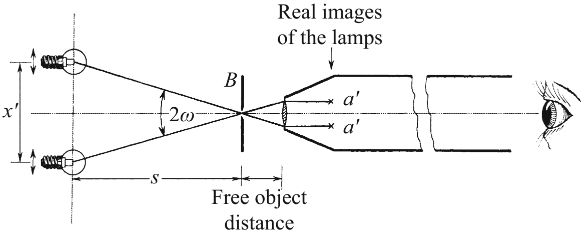


Figure 18.29 Measurement of the numerical aperture ($\sin\omega$) of a microscope objective. A very fine circular opening, illuminated by two lamps, is laid on the object table of the microscope. The lines drawn through the small opening B are the axes of very narrow light beams. The table is moved towards the objective until one sees a sharp image of the opening *with* the ocular (that is, the *free object distance* as measured from the front surface of the objective is adjusted). Then one looks into the tubus *without* the ocular and increases the distance x of the two lamps until their real images a', which lie practically in the focal plane of the objective, vanish. Now we have
$\sin\omega = \frac{x}{2}\left(s^2 + \frac{x^2}{4}\right)^{-1/2} \approx \frac{x}{2s}$.

dark, non-reflecting velvet so that the light is kept from the eye of the observer, provides *dark-field illumination*.

Because of the fundamental importance of the numerical aperture of a microscope objective, a method for measuring it is described in Fig. 18.29.

Not the arrangement of the lenses, but rather the bounding of the light beams leads us to a deeper understanding of the microscope and its resolving power. That is the essential content of this section.

"Not the arrangement of the lenses, but rather the bounding of the light beams leads us to a deeper understanding of the microscope and its resolving power".

18.13 Telescope Systems

In our description of optical instruments up to now, there was no place for a particularly simple telescope with a modest magnification and an upright image, known under the name *terrestrial telescope* or *Galilean telescope* and indispensable for mariners. For this reason, we add here a second description of telescope designs, applicable to all types.

In the usual applications of the KEPLER telescope, the object distance is very large compared to the focal length of the objective. Then the image of a distant object point lies in the focal plane of the objective. The object-side focal plane of the ocular is coplanar with it (cf. Fig. 18.26). In this manner, an optical path is obtained which is termed *telescopic* or "afocal": From an object point, a parallel, collimated light beam passes to the objective, and a parallel beam again emerges from the ocular, however with a smaller diameter. This is shown in the demonstration experiment illustrated by Fig. 18.30a for an object point lying far away on the optical axis.

To continue this demonstration experiment, we let the object point oscillate from above to below the optical axis (Fig. 18.30b). This motion allows us to discern the position of the exit pupil with great clarity, i.e. the common intersection of all the light beams on the image side. The beams maintain their parallel boundaries before and after passing through the telescope, but – and this is the decisive point – their angles of inclination relative to the optical axis are different before and after passing through the telescope. As before, in Fig. 18.26, we refer to these angles of inclination as the angles

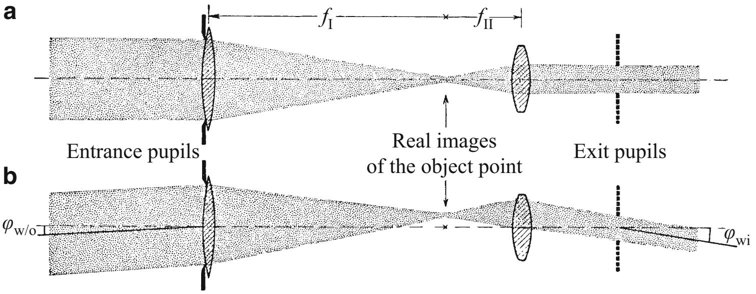


Figure 18.30 A demonstration experiment showing the 'telescopic' optical path in the KEPLER telescope for distant object points (in Part a, on the optical axis; in Part b, below it). (Threefold magnification of the angle of vision, cylindrical lenses). The vertices of the principal-ray angles of inclination $\varphi_{\mathrm{w/o}}$ and φ_{wi} lie at the centers of the entrance and exit pupils. The image is inverted. The optical setup is similar to that in Fig. 18.10. It permits us to periodically vary the angle of inclination $\varphi_{\mathrm{w/o}}$ of the parallel light beam incident from the left. The position and the formation of the exit pupil can be clearly seen. To mark the boundary rays, it is advisable to place a red filter in front of the upper rim of the entrance pupil, and a green filter in front of its lower rim.

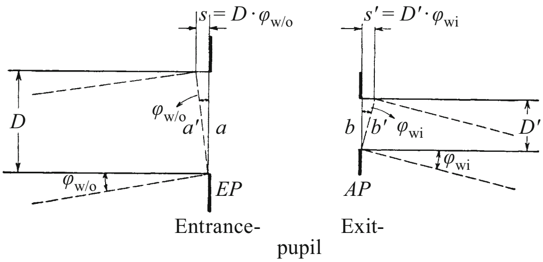


Figure 18.31 The derivation of the relation between the angular magnification and the change in the diameter D, D' of the light beam

of vision φ_{wi} and $\varphi_{\mathrm{w/o}}$ (with and without the telescope), and obtain a quantitative expression for the magnification:

$$\text{Magnification } x = \frac{\varphi_{\mathrm{wi}}}{\varphi_{\mathrm{w/o}}} = \frac{\text{Beam diameter before the telescope}}{\text{Beam diameter after the telescope}} = \frac{f_{\mathrm{I}}}{f_{\mathrm{II}}}\,. \qquad (18.6)$$

The fact which we have obtained experimentally here can be readily understood: Fig. 18.31 repeats schematically the demonstration of Fig. 18.30b, but now only the beam boundary rays before and after the telescope are drawn in. Lines a and b, perpendicular to the rays, have been added to mark the wavefronts. Then we can imagine the incident beam to be tipped by a small angle into the position shown by dashed lines. a is converted to a' and b to b'. In this process, the optical paths s and s' must remain equal (sine condition (18.2), see also Fig. 18.16). Then we have $D' \cdot \varphi_{\mathrm{wi}} = D \cdot \varphi_{\mathrm{w/o}}$. From this, as can readily be seen in Figs. 18.30 and 18.32, we obtain $\varphi_{\mathrm{wi}}/\varphi_{\mathrm{w/o}} = f_{\mathrm{I}}/f_{\mathrm{II}}$.

According to this analysis, in order to construct a telescope, we need only set up a telescopic optical path. This can be accomplished with other optical arrangements, also, for example with one converging lens and one diverging lens. We thus can build a *terrestrial telescope*, also called the GALILEAN telescope (1609, GALILEO GALILEI, 1564–1642). The demonstration experiment in Fig. 18.32 shows the path of a light beam for one distant object point on the optical axis and one point below the axis.

Knowledge of the telescopic optical path leads us to a simple procedure for measuring the *telescopic magnification* or "*power*"; we need only measure the diameter of a collimated light beam before and after it passes through the telescope and then we can apply Eq. (18.6).

The diameters of the light beams are the same as those of the entrance and the exit pupils. With a properly-constructed telescope, the entrance pupil is practically always the bezel of the objective lens. The exit pupil, the image of the objective bezel as projected by the ocular, is accessible only in the KEPLER telescope and its variants (e.g. prism binoculars). For the terrestrial telescope, it is a virtual image located inside the telescope tube,

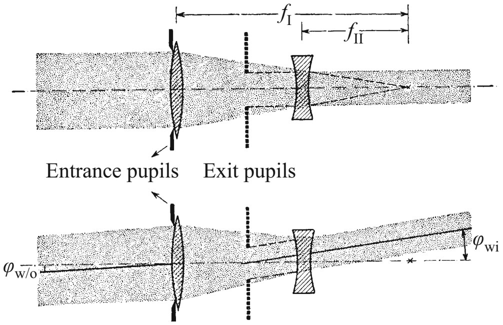


Figure 18.32 A demonstration experiment showing the 'telescopic' optical path in the GALILEAN telescope for one distant object point on the optical axis and one point below the axis (magnification of the angle of vision 2.2x). The exit pupil is a virtual image of the objective-lens bezel, projected by the ocular. Between the objective and the ocular, there is, unlike the KEPLER telescope, no image of the object point. The terrestrial telescope is constructed to give only moderate magnification (around $2-6$ x, upright image). Its main advantage is the limited number of glass surfaces and the resulting small losses in light intensity. This type of telescope is today still unmatched as a "night glass".[C18.8]

C18.8. However, among the modern "night glasses" there are also binoculars and telescopes with so-called residual light amplifiers, which make use of the infrared radiation present even on the darkest night.

between the objective and the ocular (cf. Fig. 18.32). We hold a KEPLER telescope with its objective pointed to the sky or to a bright window, and look into the ocular from a distance of around 30 cm. Then we see the exit pupil as a small, bright disk, seemingly suspended in front of the ocular. Its diameter can be measured with a millimeter ruler. The diameter of the objective lens, divided by this exit pupil diameter, yields the magnification of the telescope. With a terrestrial telescope, we must instead carry out the demonstration experiment as shown in Fig. 18.32 and determine the diameters of the beams.

18.14 The Field of View of Optical Instruments

Preliminary remark: When the unaided eye is used, often the *field of view* is limited by some sort of obstacles, for example the frame of a window. We view very small fields of view with *steady* eyes, but when the field of view encompasses several degrees or more, our eyes are *moving*: the eyes *glance*, they carry out (unconscious) jerky rotations in their sockets and *fix* individual regions within the field of view during brief pauses. These motions can be assisted by rotations and shifts of the entire head, but in that case, we see the separate regions within the field of view one after the other. That makes it difficult to maintain an overview. Looking through a keyhole is a good example.

In optical instruments, the objective and the ocular are without doubt the essential lenses. For practical construction of the instruments, however, they are not sufficient. Without additional lenses, the field of view is too restricted. The required auxiliary lenses are called *condensers* or *field lenses*. Examples are more instructive than verbose explanations of the general kind.

"Examples are more instructive than verbose explanations of the general kind".

First of all, the *projector*, with which slides, transparencies, or film, that is typical "secondary light sources", are imaged on a (usually large) screen[1]:

Figure 18.33, top part, shows an incorrectly constructed projector with a light source (arc-lamp crater), slide and the objective which forms the image. On the screen, we see only a small section from the center of the slide. The field of view is much too small (and its boundaries are not focussed). The reason: Here, the bezel of the objective acts as a *field of view aperture*. It allows only the light from a limited angular range α to pass from the lamp to the screen. The ray r has no physical significance, since no light beam is propagating in its direction. Therefore, the outer parts of the slide cannot be imaged on the screen. This can be easily corrected (Fig. 18.33, bottom): We place a large lens immediately in front of the slide, a *condenser*, and use it to image the light source onto the opening defined by the objective. Then all the light which passes through the slide will also pass through the objective.[C18.9] The slide thus appears in its entirety on the screen; its edges are sharply focussed. Now, the mounting frame of the slide acts as the field of view aperture. Its image limits the field of view as an "exit window" and has the correct orientation, i.e. in the plane of the image on the screen.

C18.9. The essential characteristic of a slide or transparency, which was presupposed here, is that it allows light to pass through without deflecting ("scattering") it. For an image which is applied to a plate of matte glass, the field of view is *not* limited by incorrect illumination, since the light is scattered in all directions. It is simply less bright. See the footnote in this section.

In general: Just as the aperture (or its image) limits the *opening angles* ω and ω' as a pupil, so do field-of-view diaphragms (or their images) limit the *angles of inclination of the principal rays* φ and φ' as *exit windows*.

The condenser must be matched to the distance between the objective and the slide. For projection with varying image sizes and distances to the screen, one requires objectives of different focal lengths. For each one of them, a matching condenser must be available.

In a precisely corresponding manner, we can use condenser lenses in a microscope and in a KEPLER telescope. As a rule, they are combined within a short tubus with the eyepiece lens. This combination is also referred to as an *ocular*.

In contrast to the usual descriptions, one seldom observes the image of a microscope or a telescope with a steady eye. Frequently, *turning of the eyeball and shifts of the entire head are necessary*. The reason is that the angular region of greatest visual acuity includes only a few degrees of arc. It is symmetric around the "fovea centralis", the

[1] For demonstration purposes, among other reasons, one can approximate a slide to a *primary light source* by putting it onto a surface which is emitting light in all directions. This could be e.g. a glass plate which is either itself fluorescent (primary light source) or a scatterer in the form of 'frosted' or 'matte' glass (secondary light source).

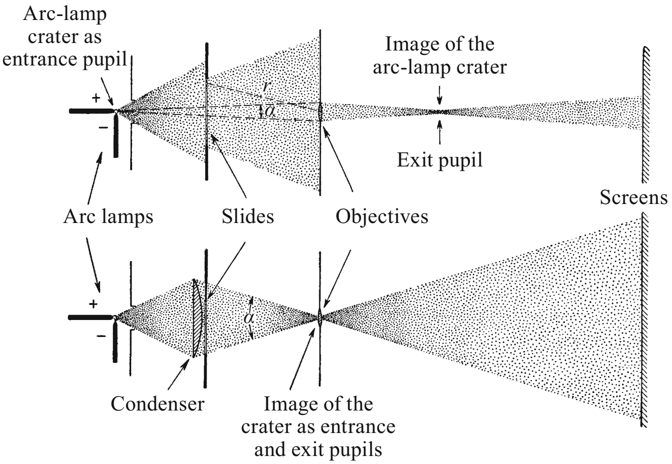

Arc-lamp crater as entrance pupil
+
−
r
α
Image of the arc-lamp crater
Exit pupil
Screens
Arc lamps
Slides
Objectives
+
−
α
Condenser
Image of the crater as entrance and exit pupils

Figure 18.33 Top: An incorrectly constructed slide projector. The objective bezel as aperture limits the angle of vision $\alpha = 2\varphi_{\max}$, i.e. the largest usable angle between two principal rays on the object side. The apex of this angle lies as always at the center of the entrance pupil (compare Fig. 18.12). – Bottom: A correctly constructed slide projector. The condenser forms an image of the light source (arc-lamp crater) at the objective (the path of a partial beam which forms the image, and its opening angle ω, can be seen in Fig. 18.12). The mount of the slide is the limiting aperture for the field of view. From its edges, principal rays with a large field-of-view angle $\alpha = 2\varphi_{\max}$ lead to the center of the entrance pupil which plays a decisive role in the image formation. In the example drawn, this entrance pupil covers only a small central spot on the objective lens which projects the image. In lecture rooms for up to 500 students, a 5-ampere arc lamp is quite sufficient. Using incandescent (filament) lamps, one can in fact use the entire diameter of the objective, but they add an unnecessary complication when projecting physical demonstrations; the same applies to condensers whose front surface is not freely accessible.

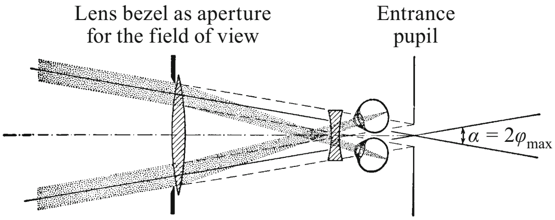

Lens bezel as aperture for the field of view
Entrance pupil
$\alpha = 2\varphi_{\max}$

Figure 18.34 At low magnifications, the exit pupil of a terrestrial telescope (Fig. 18.32) has a larger diameter than the entrance pupil of a human eye. The telescope and eye together make use of an entrance pupil which lies within the skull of the observer. Its center is, as always, the point of intersection of the object-side principal rays. The largest usable angle of inclination of the principal rays determines the angle of the field of view $\alpha = 2\varphi_{\max}$. The bezel of the objective acts as diaphragm for the field of view. When α is exceeded, the cross-section of the beam takes on the shape of a lune (a figure bounded by two circular segments of different radii). The image fades away towards its edges (vignetting).

central region of the retina with the highest density of receptors. The visual acuity decreases within ± 2° to half its maximum value and within ± 10° to only one-fifth of its maximum value. These motions of the eye and the head have to be considered when determining the field of view.

In using a KEPLER telescope, one usually moves the eye in front of the exit pupil of the telescope as if looking through a keyhole (Figs. 18.26, 18.30). With a terrestrial telescope, the eye makes use of only a portion of the objective surface for each "instantaneous observation" (action or moment of time!). This is illustrated by Fig. 18.34 for two extreme positions of the eye.

18.15 Imaging of Three-Dimensional Objects and Depth of Field

In our description of the process of image formation up to now, an image point was identified with the point where the light beam has its smallest cross-section (the 'waist'). This corresponds to the usual practice, but is by no means always generally correct. Think for example of the pinhole camera, known to every child (Fig. 18.35). It makes use of narrow light beams without any constriction on the image side. Nevertheless, it yields good images (which are completely free of distortion and are planar). This is rather surprising. The image points, i.e. the diffraction pattern of the opening, is under similar circumstances 20 times larger in a pinhole camera with a 1 mm opening than with an objective lens of 20 mm diameter (Eq. (16.23)). But an artist can also paint very satisfying pictures using broad brush strokes. This is due to psychological processes and does not belong in this section (see also Sect. 15.2). For us at this point, the often-repeated experience suffices: *High-quality images which are satisfactory for our eyes are by no means identical to images of the highest resolution*.

Even the most technically perfect lenses can form only an image of an object *plane* on an image *plane*. These two planes must be perpendicular to the optical axis. Nevertheless, in practice one often wants to image three-dimensional objects onto an image plane. As is well known, usually the resulting images are quite adequate: The eye, binoculars and cameras in general have a considerable *depth of focus* or "*depth of field*". That is however due only to a peculiarity of our eyes as mentioned above; the eye does not always treat just the narrowest constriction of a light beam as an image point.

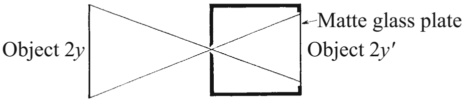


Figure 18.35 A pinhole camera

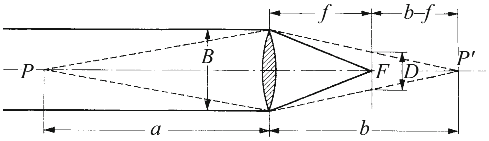


Figure 18.36 The calculation of the near-point distance a_{min} of an objective which is corrected to "infinity" (an "infinitely long" object distance). We have

$$\frac{D}{B}=\frac{b-f}{b}, \quad \frac{1}{b}=\frac{1}{f}-\frac{1}{a}.$$

The combination yields the general form of Eq. (18.7).C18.11

The objective lenses of photographic apparatus and telescopes are corrected for an "infinitely" distant object plane. Thus, object points "at infinity" are imaged as *points* on the image plane (in this case the focal plane; Fig. 18.36). At the same time, all the object points which are closer to the lens appear in the image plane not as points, but rather as small "disks". Their diameter D increases as the distance a from the object point to the lens becomes smaller. Finally, at the *near point distance* a_{min}, it reaches a limiting value D_{max} which is no longer acceptable to the eye. It is given by

$$a_{\text{min}}=\frac{fB}{D_{\text{max}}}=\frac{f^2}{N_{\text{f}}D_{\text{max}}}. \tag{18.7}$$

(N_{f}=[focal length f/lens diameter B] is the *focal ratio* of the lens.C18.10 The derivation is shown in Fig. 18.36; see also Comment C18.11.)

The near-point distance a_{min} is thus proportional to the *square* of the focal length at a given focal ratio N_{f}.

Numerical example: For an image size of 24 mm×36 mm, the empirically-determined limiting value is $D_{\text{max}} = 50\,\mu\text{m}$. At a focal ratio of $N_{\text{f}} = 5$ and a focal length of $f = 2\,\text{cm}$, we find $a_{\text{min}} = 1.6\,\text{m}$. This means that all the object points which are more than 1.6 m from the lens will be imaged simultaneously with an acuity which is satisfactory for the eye. Or put differently: The "depth of focus" in this example extends from a minimum distance of 1.6 m to infinity.

Equation (18.7) gives the physical rationale for two well-known facts:

1. Nature has evolved the eyes of the largest mammals (elephants and whales) to be not much larger than those of humans.
2. The development of photographic technology led to the production of 35 mm cameras.

C18.10. The focal ratio $N_{\text{f}} = f/B$ or *relative aperture* of a lens is also called the "f-stop", e.g. of a camera lens. Equation (18.7) is a special case (for the near-point distance a_{min}) of a general equation which applies to thin lenses. Compare Fig. 18.36 and Comment C18.11.

C18.11. The first equation, $\frac{D}{B}=\frac{b-f}{b}$, can be derived directly by comparing the two similar triangles on the right of the lens (with a common baseline $P-P'$, altitudes $B/2$ and $D/2$, and widths b and $(b-f)$); set the ratios of their altitudes and of their widths equal.

The second equation, $\frac{1}{b}=\frac{1}{f}-\frac{1}{a}$, is the "imaging formula" which also holds generally for thin lenses. It was derived in Sect. 16.7 (as Eq. (16.13)).

Multiplying this second equation by fb and reordering the terms gives $fb/a = b-f$. Inserting this into the first equation yields the general form of Eq. (18.7).

18.16 Perspective

Planar images of *three-dimensional* objects always have a certain geometrical *perspective*; that is, they exhibit a certain ratio between the size and the distance of objects which are behind one another. An artist may reproduce such a perspective by using a *central projection*. This is done as shown in Fig. 18.37: The artist places a transparent screen W between the objects and one eye, and notes the points of

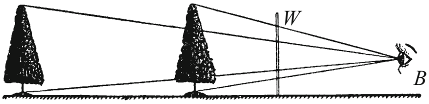


Figure 18.37 A central projection for representing three-dimensional objects on a flat image plane W (B is the eye of the artist)

intersection of the lines of sight to various objects. The artist thus employs the *pivotal point of the eye* as the center of projection.

For image formation using a lens, the lens is placed between the objects and the screen. This is also a central projection, but with two centers of projection. They lie at the centers of the entrance and the exit pupils. *Therefore, the boundaries of the light beams are decisive for perspective, also.* We will document this with an impressive demonstration experiment.

In Fig. 18.38, two brightly gleaming matte-glass windows of the same size are placed at different distances from the lens. One of these windows is in fact somewhat in front, the other somewhat behind the plane of the drawing. The back window is marked with **H**, the front window with **V**. The lens has a large diameter, but a narrow aperture and a thin light beam are used. As a result, both windows appear equally sharp on the screen, adjacent to each other. During the experiment, the setup remains unchanged (Fig. 18.38a); only the aperture is moved along the optical axis. The experiment is carried out in three steps:

1. The aperture is set up immediately in front of the lens (Fig. 18.38b). Both pupils are practically contiguous with the center of the lens; it serves as the center of projection. The more distant window **H** appears smaller on the screen than the closer window **V**.

2. The aperture is shifted to the image-side focal point F' (Fig. 18.38c). This moves the object-side center of projection (the center of the entrance pupil) to the left towards 'infinity': The two images of **H** and **V** have now become the same size on the screen.

> Figure 18.38c shows the limiting case of an object-side telecentric optical path. This is often used; it is for example indispensable in a measuring microscope.

3. In Fig. 18.38d, the aperture is shifted on the image side to beyond the focal point F'. Then the object-side center of projection (the center of the entrance pupil) moves closer to the window **H** than to the window **V**. The result is that the image of **H** becomes larger (!) on the screen than the image of **V**; the perspective is inverted.

We can thus vary the geometrical perspective of an image over a large range simply by shifting an aperture which limits the boundaries of a light beam. So much for the demonstration experiment.

Pictures painted by an artist are supposed to be viewed from the same center of projection as that used by the artist. One should use only

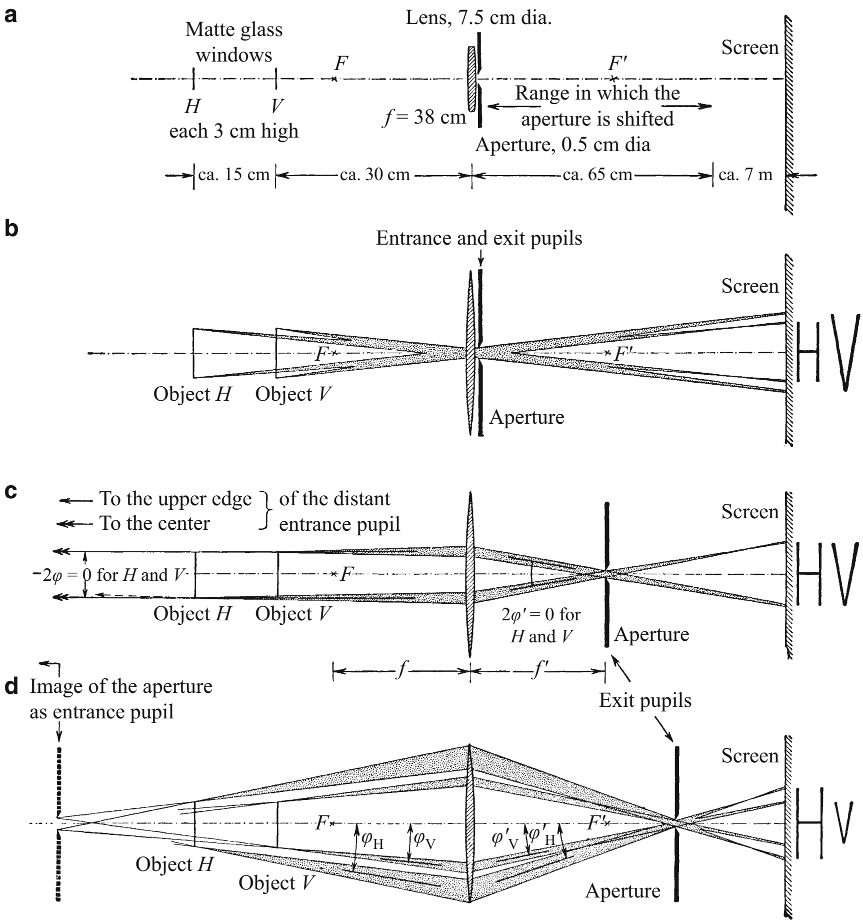


Figure 18.38 The influence of the light-beam boundaries on perspective. Part a: The experimental setup. One of the windows (**V**) is to be thought of as being some distance in front of the plane of the page, the other (**H**) somewhat behind it. Parts b–d: The size ratio between **H** and **V** is changed by shifting the aperture which limits the light beam boundaries. On the object side, the center of the entrance pupil always serves as the center of projection. The lens "looks at" the objects **H** and **V** from this point. In Part c, for clarity only the beams coming from the top of **V** and from the bottom of **H** are drawn. At intermediate positions between b and c, the entrance pupil lies as a virtual image to the right of the aperture **(Video 18.4)**. – A good free-hand attempt at the inverted perspective in Part d: Hold a lens of around 10 cm diameter and 20 cm focal length (a reading glass) about 30 cm in front of your eye and observe a box of matches. You will see the more distant edge of the box larger than the closer edge.

one eye and place it at the position *B* in Fig. 18.37. Then, with high-quality paintings, one has the impression of seeing a natural three-dimensional view.

In a photographic camera, the principal rays propagate from the center of the exit pupil to the film or detector. The center of the exit pupil serves as the image-side center of projection. *Therefore, when looking at a photo, the pivot point of the eye should be placed at this center of projection.* This presents no difficulties: In the objective

Video 18.4: "Perspective"
http://tiny.cc/5eggoy
The two letters **H** and **V** are somewhat fuzzy due to the limited depth of focus (Sect. 18.15)! Note also the 'free-hand experiment' described in the caption of Fig. 18.38.

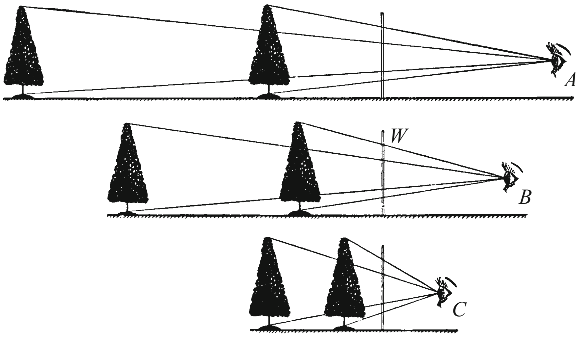


Figure 18.39 Objects of the same size at different depths are projected from the centers A, B, and C onto the same image plane W. The points of intersection of the lines of sight with the image plane are the same for all three examples shown here. – With these figures, we indicate the distortion of perspective which results when a picture is viewed from the incorrect distance: An image cast from the center B appears foreshortened when seen from C, and stretched out when viewed from A.

lenses common today, the entrance and the exit pupils nearly coincide with the center of the lens. One thus needs only one center of projection, as in the schematic of Fig. 18.38b. Furthermore, the film is always placed near the focal plane of the objective. Then we find the following rule:[C18.12] *One should always view a photograph with one eye and at a distance between the pivot point of the eyeball and the photo which is equal to the focal length of the camera lens that was used to take it.* For focal lengths of around 25 cm and upwards, this can be readily done. However, the usual 35 mm cameras have considerably shorter focal lengths (not to mention the electronic cameras in smartphones, tablet computers etc.), at most a few centimeter. In this case, one has to use a magnifying lens between the photo and the eye; then the correct distance can be maintained. When this rule is observed, every photo shows a surprisingly good sculptural effect and lifelike perspective.

C18.12. We make special mention of the rule described here for viewing photos. As Pohl says further on, we have forgotten how to see the three-dimensional structure of photos, and treat them as if they were only two-dimensional images.

Good-quality magnifying lenses should be designed for the *gazing* eye, and the distance between the pivot point of the eyeball and the lens should be fixed by a suitable shape of the lens mount (bezel). For an x-fold linear magnification of the photographic print relative to the negative, the eye-to-photo distance should be equal to xf. Unfortunately, this condition can be fulfilled for only a few members of the audience in a large motion-picture theater, and the seats where it is met vary with the magnification of the film.

All pictures when seen with one eye, both those painted by artists and those taken by a camera, should give a *three-dimensional* impression, even when viewed from the incorrect distance, although the perspective may then be distorted. The depth of field appears too shallow when the viewing distance is too short, and too deep when it is too long (Fig. 18.39). But today, our visual senses have all been dulled by the flood of images in magazines, television and digital media.

We have given up on seeing pictures as three-dimensional and experience them only as *flat surfaces*. Only when the circumstances are unusual does the true ability of our eyes again emerge. For example, we see the *two-dimensional* images in the focal plane of a telescope as viewed through the ocular lens as *three-dimensional*, but the depth of field of all the objects is foreshortened. The long axis of a street or boulevard is especially suitable for seeing this effect; the image is projected by an objective with a long focal length f and should therefore be viewed with the correct depth of field from this same distance f. An ocular lens with a focal length of f would however make the magnification of the angle of vision equal to one, and thus would defeat the purpose of the telescope. Only with an ocular lens of *short* focal length can the angles of vision be increased and the image thus magnified. But then it makes the viewing distance unavoidably too short, so that we see all depths in the image as foreshortened.

Still more impressive is a reversal of this experiment: We look through the telescope 'backwards' and use its objective as an ocular. Then we see the depth of field stretched out in a very comical manner. The ocular, with its short focal length, projects a two-dimensional image, and we view it through the objective from much too great a distance.

More details on image formation, in particular the production of visible images of invisible objects, can be found in Sect. 21.11.

Exercises

18.1 Figure 18.1 shows the graphical construction of the image point P' which belongs to the object point P. Use this drawing to derive the lens formula $1/a + 1/b = 1/f'$ (Eq. (16.13)) (a = the object – lens distance, b = the image – lens distance). (Sect. 18.2; see also **Exercise 16.3**).

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_18) contains supplementary material, which is available to authorized users.

Radiation Energy and Beam Limitation

19

19.1 Preliminary Remark

In our whole discussion of image formation and optical instruments, the details of lens fabrication and ray diagrams have not been emphasized; instead, we have used the *limiting boundaries of light beams* as an explanatory principle. This decisive point will also lead us to an understanding of energy transport by radiation, whether or not it is accompanied by image formation.

19.2 Radiation and Opening Angle: Definitions.[C19.1] LAMBERT's Cosine Law

Up to now, we have always treated the "image points" in an optical image as small regions or surface elements, corresponding to experimental fact; but we have presumed the object points without comment to be *mathematical points*. That has caused us no problems thus far, but we should be careful to correct it. In reality, radiation of non-zero energy is always emitted from a finite surface element dA.[C19.2]

In Fig. 19.1, at the left, we show dA as a small glowing piece of metal with a *matte black* surface. It acts as an *emitter*. Its front surface emits radiation in all directions within a hemisphere, so that within the time interval dt, it emits the energy dW. How is this energy distributed in space? In order to answer this question, we detect the radiation with a radiometer (Sect. 15.3). It serves as a small *receiver*. Let its free surface area be dA'; this surface is oriented perpendicular to the direction of propagation of the radiation ("normal incidence"). Furthermore, both the dimensions of the emitter, dA, as well as those of the receiver, dA', are chosen to be small compared to their mutual distance R.

The signal detected by the radiometer (i.e. its deflection) corresponds to the *radiant flux* d$\dot{W}$ which falls on the receiver, that is energy/time with the unit 1 watt, also called the energy current. We now vary the

C19.1. Some of the quantities discussed here, which are related to energy transport by radiation, were already introduced in Chap. 15. A detailed treatment of the analogous quantities which are defined especially for *light*, taking the perception of brightness by the human eye into account, will be given in Chap. 29.

C19.2. The *infinitesimal* quantities d$\dot{W}_\vartheta$, dA, dA' and dΩ in this chapter represent small but nonzero values. The surface elements dA and dA' are chosen to be so small that the overall shape of the surface plays no role. The radiant flux d$\dot{W}_\theta$ propagates along the ϑ direction for example from the emitter area dA to the receiver area dA'. In some cases (e.g. the definition of the radiance L_e), a true differential quotient is meant (second derivative of d$\dot{W}_\theta$ w.r.t. dA and dΩ). For a general exposition of radiometric quantities, see e.g. the article by Ian Ashdown, www.helios32.com/Measuring%20Light.pdf

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_19

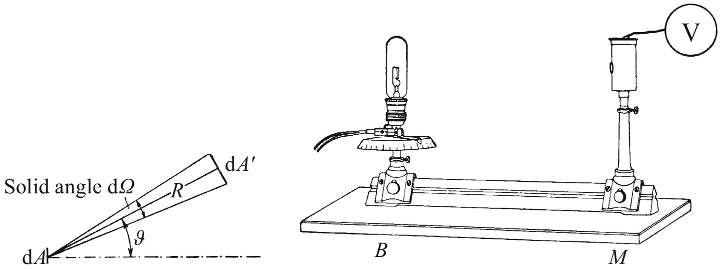
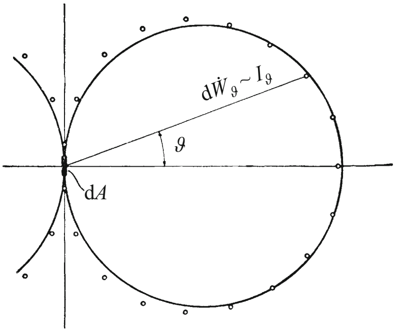


Figure 19.1 Measuring the radiant flux $\mathrm{d}\dot{W}$ which is emitted from a surface element dA, for example from a tungsten-ribbon lamp (B), under different angles of inclination ϑ into the solid angle dΩ (dA' is the surface area of a radiometer, e.g. a thermopile (M)). At the left is a schematic of the emission geometry, at the right the experimental arrangement. The angle ϑ is varied here only within the horizontal plane.

Figure 19.2 The angular distribution of the radiant flux reaching a detector of area dA'. (The points were measured as shown in Fig. 19.1, right). The large circles are calculated from Eq. (19.1) (Lambert's cosine law).



size of dA, dA', R and ϑ, and find (where we denote the proportionality factor by L_e):[C19.3]

C19.3. In general, ϑ is the angle between the surface normal vector of dA and the vector R, whereby R can range over the whole hemisphere in front of dA. In the following examples, R is however often allowed to rotate only within the horizontal plane, as in Fig. 19.1.

$$\mathrm{d}\dot{W}_\vartheta = L_\mathrm{e} \cdot \mathrm{d}A \cdot \cos\vartheta \cdot \frac{\mathrm{d}A'}{R^2}\,. \tag{19.1}$$

The influence of the quantities dA, dA' and R is as expected from simple geometrical considerations. *The proportionality of the radiant flux to* $\cos\vartheta$ *in the direction* ϑ*, in contrast, can be derived only from experiments.* It is generally obeyed only approximately. (It is called Lambert's cosine law, described in 1760 by Johann Heinrich Lambert, 1728–1777). An example is shown in Fig. 19.2. Lambert's law is however exact for a small opening dA which is the source of the emission through the wall of a cavity at a uniform temperature, a "black body radiator" (Sect. 28.4).

In Eq. (19.1), which was found empirically, the ratio dA'/R^2 refers to the solid angle[1] dΩ. It is an open cone. Its apex is at the center of

[1] The unit of a solid angle is, like the unit of every angle, simply the number 1. It is often expedient to give the number 1 in this connection the name 'steradian' (sr). See also the footnote at the end of Chap. 17. More details are given in Vol. 1, Sect. 1.5.

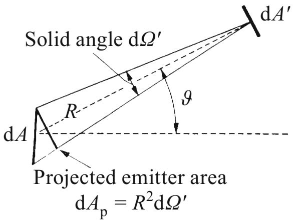


Figure 19.3 The "projected emitter area" $\mathrm{d}A_\mathrm{p} = R^2 \mathrm{d}\Omega'$. Here, $\mathrm{d}\Omega'$ is the solid angle under which an arbitrarily-shaped emitter is seen from the position of the receiver. In the special case of a planar emitter $\mathrm{d}A$ as sketched here, the projected emitter area is $\mathrm{d}A_\mathrm{p} = R^2\mathrm{d}\Omega' = \mathrm{d}A \cos\vartheta$.

the surface element $\mathrm{d}A$, and thus at the center of the emitting surface. Its base is the irradiated surface element $\mathrm{d}A'$, and thus the area of the receiver. Furthermore, $\mathrm{d}A \cos\vartheta = \mathrm{d}A_\mathrm{p} = R^2\mathrm{d}\Omega'$ can be seen from Fig. 19.3 to be the *projected* or *apparent emitter area*, and $\mathrm{d}\Omega'$ is the solid angle under which the emitter is seen from the location of the receiver.

The proportionality factor L_e in Eq. (19.1) (in differential form) is then

$$L_\mathrm{e} = \frac{\mathrm{d}^2 \dot{W}_\vartheta}{\mathrm{d}\Omega\, \mathrm{d}A_\mathrm{p}}\,;$$

it characterizes the emitter. The quantity L_e is called the *radiance* of the emitter. Its unit is 1 watt/(steradian $\cdot$m^2) = 1 watt/m^2.

The experimental introduction of the radiance L_e is by no means limited to the special case of a *planar* emitter for which LAMBERT's cosine law holds. Imagine for example in Figs. 19.1 and 19.2 that the emitter is a glowing cylinder which is perpendicular to the plane of the page. Then the radiant flux $\mathrm{d}\dot{W}$ that it emits is independent of ϑ. The angle ϑ is varied only within the horizontal plane (as in Fig. 19.1, right). Instead of Fig. 19.2, we would then have a circle with the emitter at its center. We would find $\mathrm{d}\dot{W} = L_\mathrm{e}\mathrm{d}\Omega\mathrm{d}A$, and thus the proportionality factor (the radiance) would be $L_\mathrm{e} = \mathrm{d}^2\dot{W}/\mathrm{d}\Omega\mathrm{d}A$.

The quantity

$$I_\vartheta = \frac{\mathrm{d}\dot{W}_\vartheta}{\mathrm{d}\Omega} = \frac{\text{Radiant flux in the direction } \vartheta}{\text{Solid angle}} \qquad (19.2)$$

or

$$I_\vartheta = L_\mathrm{e} \cdot \mathrm{d}A_\mathrm{p} = \text{Radiance times projected emitter area} \qquad (19.3)$$

characterizes the radiation of the emitter in the direction ϑ, and therefore, I_ϑ is called the *radiant intensity* in the direction ϑ. Its unit is 1 watt/steradian (W/sr).

The same radiant intensity I_ϑ can be produced by emitters of very different sizes. At white heat, a small surface area is sufficient; at a dull red glow, a large area would be necessary.

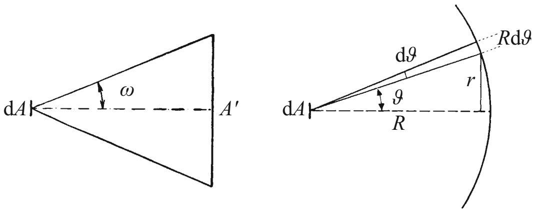


Figure 19.4 At the left: The calculation of the radiant flux $\mathrm{d}\dot{W}_\omega$ emitted from $\mathrm{d}A$ (emitter) and transmitted to A' (receiver) according to Eq. (19.5). The small emitter $\mathrm{d}A$ which is located at a distance R from the receiver A' emits radiation with the radiance L_e. At the right: a spherical surface construction to aid in the calculation.

The receiver, the small irradiated surface element $\mathrm{d}A' = \mathrm{d}\Omega R^2$, registers the radiant flux $\mathrm{d}\dot{W}$ arriving *perpendicular* to its area. The quotient

$$E_\mathrm{e} = \frac{\mathrm{d}\dot{W}}{\mathrm{d}A'} = \frac{\text{Incident radiant flux}}{\text{Receiver area}} = \frac{\text{Radiant intensity } I_\vartheta \text{ of the emitter}}{(\text{Distance } R \text{ to the emitter})^2} whichgoesfromtheemitter \tag{19.4}$$

is given the name *irradiance* or *irradiation intensity*. Its unit is 1 watt/m^2.[C19.4]

C19.4. The irradiance is thus an *energy flux density*, i.e. the energy per time interval and surface area arriving at a receiver. It is denoted in the literature by the letter E or E_e.

Thus far, we have assumed that the receiver $\mathrm{d}A'$ is small compared to the distance R; the surface element $\mathrm{d}A'$ was supposed to be practically perpendicular to the radiation direction. We now relax this limitation; however, the emitter is initially still supposed to have a small area $\mathrm{d}A$. In Fig. 19.4, at the left, a large circular area A' is irradiated with the opening angle ω, and, apart from its center, it receives the radiation at an angle (i.e. *not* at normal incidence). Then this receiver A', presuming the validity of LAMBERT's cosine law, receives the radiant flux

$$\mathrm{d}\dot{W}_\omega = \pi L_\mathrm{e} \mathrm{d}A \sin^2\omega \tag{19.5}$$

(or $I_\omega = \mathrm{d}\dot{W}_\omega/\mathrm{d}\Omega = L_\mathrm{e}\cdot\mathrm{d}A$). It is sent out by the emitter, which has an area $\mathrm{d}A$ and a radiance L_e.

Derivation: To calculate the radiant flux which is received by the receiver area A', we use a construction as in Fig. 19.4 (right), a spherical surface in front of the receiver A'. All of the radiation which reaches A' must first pass through this virtual sphere. We decompose its surface into a series of narrow circular zones which are ring-shaped and concentric around the direction vector $\boldsymbol{R}$ (i.e. the $\vartheta = 0$ axis, dashed in the figure). Each has an area of

$$\mathrm{d}A'_\text{ring} = 2\pi r \cdot R\,\mathrm{d}\vartheta = 2\pi R^2 \sin\vartheta \cdot \mathrm{d}\vartheta\,.$$

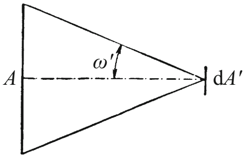


Figure 19.5 A large emitter A which emits at the radiance L_e irradiates a small receiver dA' (Eq. (19.8)). For this light beam, we cannot draw any sort of *simple* wave picture.

Each of these ring-shaped circular zones receives a radiant flux (from Eq. (19.1)) equal to

$$d\dot{W}_\vartheta = L_e dA \cos\vartheta \frac{dA'_\text{ring}}{R^2} = 2\pi L_e dA \sin\vartheta \cos\vartheta \, d\vartheta = 2\pi L_e dA \sin\vartheta \, d(\sin\vartheta)\,.$$

Integration of this flux over angles between $\vartheta = 0$ and the full opening angle $\vartheta = \omega$ yields the overall power $d\dot{W}_\omega$ which arrives at the circular receiver area A' (i.e. Eq. (19.5); cf. Fig. 19.4).

The radiant flux in Fig. 19.4 which is radiated out from the emitter dA and is incident on the circular area A' of the receiver attains its maximum value $d\dot{W}_\text{max}$ for the limiting case of $\omega = 90°$. This value is used to define the *radiant exitance* of the emitter by the equation

$$M_e = \frac{d\dot{W}_\text{max}}{dA} = \frac{\text{Radiant flux emitted to one side}}{\text{Emitter area}}\,. \qquad (19.6)$$

When LAMBERT's cosine law holds, we find from Eq. (19.5) the radiant exitance of the emitter to be:

$$M_e = \frac{d\dot{W}_\text{max}}{dA} = \pi L_e\,. \qquad (19.7)$$

When emission to both sides is considered, a factor of 2 must be included.

If we reverse the direction of light propagation (Fig. 19.5), the large area A acts as emitter and the small area dA' as receiver. We again denote the radiance of the emitter as L_e. Then the radiant flux arriving at dA' is

$$d\dot{W}_{\omega'} = \pi L_e dA' \sin^2\omega' \qquad (19.8)$$

(Derivation as above) .

Equation (19.8), i.e. the dependence of the flux $d\dot{W}_{\omega'}$ on the opening angle ω', can be illustrated by a demonstration experiment. We use a *secondary source* as emitter, e.g. a circular area on a high-quality matte white projection screen which is irradiated by an arc lamp (compare Sect. 26.10 and Fig. 26.14). Then, for varying opening angles ω', we measure the flux $d\dot{W}_{\omega'}$. ω' can be varied in two different ways, namely by changing the diameter of the circular area, or by changing the distance between the emitter and the receiver. In Sect. 19.3, an application of this important equation (19.8) will be discussed.

19.3 Radiation from the Surface of the Sun

The sun irradiates the surface of the earth at perpendicular incidence and (neglecting absorption losses in the atmosphere) with an irradiance of

$$E_\mathrm{e} = 1.367\,\frac{\mathrm{kW}}{\mathrm{m}^2}\,.$$

(Astronomers call this irradiance the *solar constant*.)

The sun's disk has an angular diameter of 32 minutes of arc as seen from the earth. Therefore, the opening angle ω' in Fig. 19.5 is equal to 16 minutes of arc, and we have $\sin\omega' = 4.7\cdot 10^{-3}$. We insert these values of the irradiance $E_\mathrm{e} = \mathrm{d}\dot{W}/\mathrm{d}A'$ and of $\sin\omega'$ into Eq. (19.8) and compute the radiant exitance M_e averaged over the surface of the sun[2]:

$$\pi L_\mathrm{e} = 6.3\cdot 10^4\,\frac{\mathrm{kW}}{\mathrm{m}^2}\,.$$

For comparison: $25\,\mathrm{m}^2$ of the sun's surface delivers about the same power as a large AC turbogenerator with a rated power of $1.5\cdot 10^6$ kW.

19.4 The Radiance L_e and the Irradiance E_e in Image Formation

In numerous cases, between the light source (the emitter) and the irradiated surface (the receiver), there is a lens or a series of lenses. With these lenses, or with any kind of imaging apparatus, one can change only the irradiance E_e at the receiver, but never the available radiance L_e. The latter is a characteristic quantity which describes the emitter. *An image of the emitter can never radiate at a higher radiance than the emitter itself.* The usable value of the radiance can in the most favorable case (absorption-free lenses or mirrors) only be conserved in an imaging process. This result, which can also be derived thermodynamically from the Second Law, will be discussed in more detail in the following.

In Fig. 19.6, a lens projects an image $\mathrm{d}A'$ of an emitter $\mathrm{d}A$. The power in this image is collected by a receiver of area $\mathrm{d}A'$. From the schematic in Fig. 19.4, we find the radiant flux

$$\mathrm{d}\dot{W}_\mathrm{wi} = \pi L_\mathrm{e}\mathrm{d}A\,\sin^2\omega_\mathrm{wi} \tag{19.9}$$

(The index 'wi' means *with a lens*)

[2] The radiation which is emitted by the sun comes from a layer about 200 km thick, which is increasingly cooler towards the outside. Near the edges of the solar disk, the paths of the radiation through the cooler regions of the layer become longer. At the very edge, one therefore measures (of course depending on the wavelength) around 60 % lower values of the radiance than at the center of the solar disk.

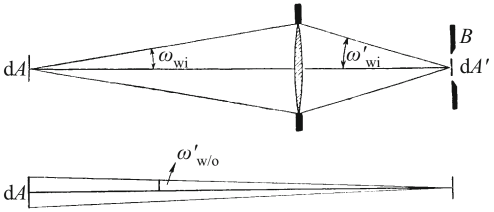


Figure 19.6 Irradiation of the receiver dA' with and without a lens. The lens increases the opening angle ω'.

which goes from the emitter dA to the lens, passes through it and produces the image dA'. Here, the lens itself acts as an emitter with a still-unknown radiance $L_{e,x}$. Its exit pupil sends out the radiant flux$^{C19.5}$

$$\mathrm{d}\dot{W}_{wi} = \pi L_{e,x} \mathrm{d}A' \sin^2 \omega'_{wi} \qquad (19.10)$$

C19.5. Since the radiant flux reaches the lens with a distribution corresponding to LAMBERT's cosine law, the lens surface on its other side again radiates with this same distribution, only that – owing to refraction – the directions are changed and the radiation is again focussed. The result is that the radiation emitted from the lens can be described by a constant radiance $L_{e,x}$.

onto the image area dA', as in the schematic in Fig. 19.5. Here, we have implicitly idealized a limiting case: We have neglected radiation losses by reflection at the lens surfaces and by absorption in the glass of the lens, and have taken the radiant flux to be the same in front of and behind the lens. In this limiting case, we can combine the two equations (19.9) and (19.10) to obtain

$$L_e \mathrm{d}A \sin^2 \omega_{wi} = L_{e,x} \mathrm{d}A' \sin^2 \omega'_{wi} \,. \qquad (19.11)$$

We make use of light beams with a wide opening angle for the imaging of dA onto dA'. Therefore, the *sine condition* (18.2) must be fulfilled:

$$\mathrm{d}A \cdot \sin^2 \omega_{wi} = \mathrm{d}A' \sin^2 \omega'_{wi} \,. \qquad (19.12)$$

Equations (19.11) and (19.12)$^{C19.6}$, when combined, yield $L_{e,x} = L_e$, an important result: *For the image* dA'*, the disk of the lens radiates with the same radiance* L_e *as the surface of the emitter; both surfaces appear to have the same "brightness"* (see Sect. 29.7). This fact will first be verified in a demonstration experiment (Fig. 19.7).

C19.6. Since dA and dA' are areas, the sine functions occur here as *squares*.

What, however, changes energetically as a result of image formation? It is the *irradiance* E_e. In Fig. 19.6, top, the lens (with a sufficient diameter) can irradiate the receiver with a larger *opening angle* ω'_{wi} and thus produce a higher radiant intensity at the position of the image than would be possible for the emitter dA *without* a lens (think of a "burning glass" !).

We use Eq. (19.8) to calculate the irradiance at the receiver in both cases in Fig. 19.6, that is the quantity

$$E_e = \frac{\mathrm{d}\dot{W}}{\mathrm{d}A'} = \pi L_e \sin^2 \omega' \,. \qquad (19.13)$$

With the lens, we have to set $\omega' = \omega'_{wi}$; without the lens, $\omega' = \omega'_{w/o}$. Then we obtain the ratio of the two irradiances with and without

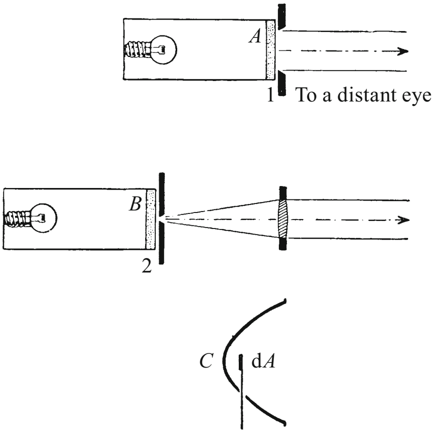


Figure 19.7 A comparison of the radiance of a small emitter with the radiance of the large area of a lens which forms its image (top and middle figures): Two similar emitters *A* and *B* consist of two identical frosted-glass disks which are illuminated from behind. After their similarity has been verified, we use two circular iris diaphragms 1 and 2 to limit the diameter of *A* to 10 cm and of *B* to 5 mm, and place *B* in the focal plane of a lens of 10 cm diameter. We then observe the emitters from a great distance and see the large *lens area* radiating with the same radiance as the *emitter A* (both appear to be equally "bright" (see Sect. 29.7)). The focal length *f* of the lens is not important. The longer *f* is, the smaller the angular region from which the radiation that emerges from the lens surface originates (this indeed reduces the radiant *flux* which passes through the lens, but the *radiance* (surface density) at the lens surface remains constant). An analogous experiment with a mirror instead of a lens is shown in the bottom figure. In order to make small glowing surfaces d*A*, for example from phosphorescent materials, clearly visible to a large audience, we put them at the focal point of an automobile headlight reflector (*C*). Then the large opening of the parabolic headlight mirror radiates with the same radiance as the small area d*A*. In spite of its triviality, this experiment often surprises even professionals.

a lens[3]

$$\frac{E_{\mathrm{e,wi}}}{E_{\mathrm{e,w/o}}} = \frac{\sin^2 \omega'_{\mathrm{wi}}}{\sin^2 \omega'_{\mathrm{w/o}}} \,. \tag{19.14}$$

The sun radiates with a radiant exitance of $\pi L_\mathrm{e} = 6.3 \cdot 10^4$ kilowatt/m^2 (Sect.19.3). Due to its great distance from the earth ($R = 1.5 \cdot 10^{11}$ m), the earth is irradiated with the very small opening angle $\omega'_{\mathrm{w/o}} =$ 16 minutes of arc (that is $\sin \omega'_{\mathrm{w/o}} = 4.7 \cdot 10^{-3}$). Therefore, for a surface element dA' on the surface of the earth, the irradiance is only $E_{\mathrm{e,w/o}} = 1.37\,\mathrm{kW/m^2}$ (for normal incidence and neglecting the losses of about 50 % in the atmosphere). With lenses or concave mirrors, we can produce opening angles ω'_{wi} of up to about 50° (that is $\sin \omega'_{\mathrm{wi}} = 0.77$). As a result, we find from Eq. (19.14) for the

[3] Here, as always, we assume the same material in front of and behind the lens, for example air.

irradiance of the solar image:

$$E_{\mathrm{e,wi}} = 1.37\,\frac{\mathrm{kW}}{\mathrm{m}^2}\left(\frac{0.77}{4.7\cdot 10^{-3}}\right)^2 = 3.7\cdot 10^4\,\frac{\mathrm{kW}}{\mathrm{m}^2}\,.$$

In order to attain a similar irradiance *without* a lens or a concave mirror, we would have to approach the sun so closely that the solar disk would reach from the horizon up to 10° past the zenith!

With a focal length of 1 m, we can project a solar image with an area of 0.6 cm^2. With the opening angle $\omega'_{\mathrm{wi}} = 50°$, the radiant flux in the solar image[4] is thus 0.6 cm^2 · 3.7 kW/cm^2 $\approx$ 2 kW. This power is the same as the power of an electric arc lamp carrying a current of 40 A at a voltage of 50 V.

Part II

19.5 Emitters with Direction-Independent Radiant Flux

LAMBERT's cosine law (Eq. (19.1)) is, as we have emphasized, an empirical limiting case. It holds exactly for a small opening into a "black-body" radiator, as we mentioned above. Planar, matte black surfaces with strong scattering or diffuse reflection are a good approximation to a black body, no matter whether their radiation is excited thermally or by some other means, for example as fluorescence.

A black body and a planar, matte surface have a common property: For both, the extinction constant, i.e. the ratio of the radiant flux that is not reflected or re-emitted, to the incident radiant flux, is independent of the angle of incidence.

A very different limiting-case law is found for the *radiation from the interior of a flat, transparent body*. For the *radiant flux* in the ϑ direction, we obtain

$$\mathrm{d}\dot{W}_\vartheta = L_\mathrm{e}\mathrm{d}A\frac{\mathrm{d}A'}{R^2} = L_\mathrm{e}\mathrm{d}A\cdot\mathrm{d}\Omega\,,$$

i.e. the *radiant intensity* in the ϑ direction,

$$I_\vartheta = \frac{\mathrm{d}\dot{W}_\vartheta}{\mathrm{d}\Omega} = L_\mathrm{e}\mathrm{d}A\,, \tag{19.15}$$

is *independent* of the angle of emission ϑ. A graph of the emitted radiant flux (Fig. 19.8) shows *one* circle with the emitter dA at its center and not, as in LAMBERT's cosine law, two circles located symmetrically to either side of the emitter (Fig. 19.2). This

[4] E. W. TSCHIRNHAUS, 1651–1708, mathematician, owner of a farm at Kieslingswalde, near Görlitz, and member of the Paris Academy from 1682 on, constructed a 'burning mirror' in 1686 with an opening diameter of 2 m and a focal length of 1.3 m, made of polished copper, and used it to *melt* materials.

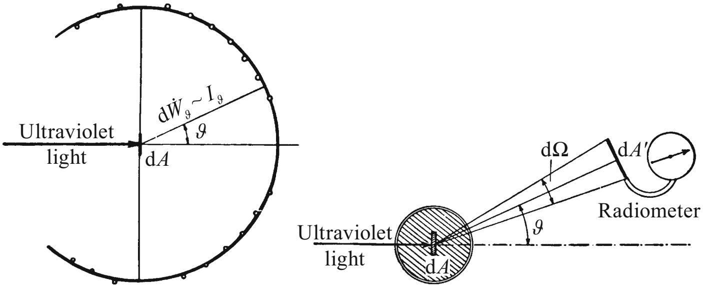
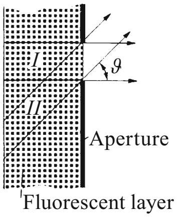


Figure 19.8 Demonstration of a directionally-independent emission intensity. The emitter is a sheet of uranium glass which is excited to fluorescence in the visible range by strongly-absorbed ultraviolet light. (To prevent reflections at the surfaces of the sheet, it is immersed in a mixture of benzene and carbon disulfide, whose index of refraction for the fluorescent light is the same as that of the glass)

Figure 19.9 The production of a radiant intensity which is independent of the emission direction ϑ: In the rectangle I and in the rhombus II, there are equal numbers of fluorescent molecules



limiting case of direction-independent radiant intensity can be implemented in various ways for a planar emitter, most simply with fluorescence radiation from a clear glass sheet. The right-hand image in Fig. 19.8 shows a suitable arrangement which minimizes disturbing reflections.

Figure 19.9 illustrates how, with this arrangement, the radiant intensity becomes independent of ϑ: Perpendicular to the emitter sheet ($\vartheta = 0$), the volume I acts as emitter, and under the angle of inclination ϑ, the volume II is the emitter. The two volumes are of equal size, and thus contain an equal number of independently emitting molecules, indicated as dots in the figure. Their emitted radiant fluxes simply add, since the glass sheet is completely transparent to the fluorescence radiation.

This independence of the direction of the radiant intensity I_ϑ has an important consequence: The radiance of the planar emitter surface, that is the quantity

$$\frac{\text{Radiant intensity } I_\vartheta}{\text{Projected emitter area } \mathrm{d}A \cos\vartheta} = L_\mathrm{e}\,,$$

is not constant, as when LAMBERT's cosine law holds; instead, L_e *increases* with increasing angle of emission ϑ: If we look from a grazing angle at the surface of the emitter sheet, we see the thin fluorescent layer with an almost dazzling luminous density.[C19.7]

C19.7. For the concept of "luminous density" (now called *luminance* L_L), see Sect. 29.3 (second footnote), and also Eq. (29.12) and Table 29.2.

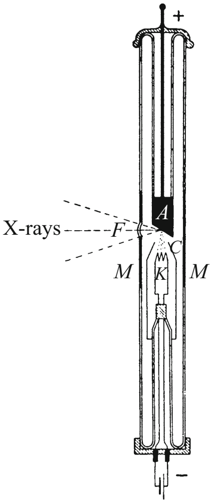


Figure 19.10 One of the many possible forms of an X-ray source with a hot cathode. The cone *C* concentrates the electrons onto a small spot on the anticathode *A*. (*M* is a metal tube, and *F* is a glass radiation window)

A direction-independent radiant intensity can also be found at the planar anticathode of X-ray tubes (Fig. 19.10). The reason: the accelerated electrons can penetrate only into a thin surface layer of the anticathode, while the X-rays themselves, in contrast, can pass through the material almost unimpeded. A practical application: One can use the X-rays which are emitted nearly parallel to the surface of the anticathode to obtain a sharply-defined focal spot of high radiance (a "streak focal line") through perspective foreshortening (W.C. Röntgen, 1896).

19.6 Parallel Light Beams as an Unattainable Limiting Case

According to all experimental evidence, "parallel light beams" or *collimated beams* can be only approximately obtained in practice. The reasons for this are already well known:

1. Every light source has a finite extension, however small it may be. Such a source can emit only beams with a nonzero opening angle ω, no matter which of all possible arrangements of pupils and lenses is used.

2. Every light beam exceeds its geometrically-constructed boundaries due to *diffraction*.C19.8

We can now add to this list: A light beam with mathematically strictly parallel boundaries would have an opening angle of $\omega = 0°$. As a result, its radiant flux as calculated from Eq. (19.5) would be zero.

For all these reasons, we should, strictly speaking, refer only to *quasi-parallel* light beams when discussing experiments.

C19.8. Even the light beams from high-quality lasers have a very small but finite opening angle owing to diffraction at their exit pupils. For example, the laser beam which was used in 1969 to measure the distance from the earth to the moon had a diameter on the lunar surface of around 1.6 km.

Interference 20

20.1 Preliminary Remark

In Volume 1, interference is treated in detail within the framework of the general topic of waves (Chap. 12). There, we considered two conditions, which can usually be fulfilled to a good approximation:

1. *Pointlike wave centers*, i.e. their diameters must be small compared to the wavelength of the radiation.

2. *Wave trains of unlimited length and a single frequency.* Only with such wave trains can we produce interferences between two *independent* emitters with the same frequency, e.g. two whistles.

If these two conditions are not sufficiently well fulfilled, we can obtain clear-cut, spatially fixed interference patterns only by taking special measures. This is the case for light, in particular; that is why we have postponed the treatment of such measures to the section on optics.

Wave trains of limited length are generically called *wave groups*. They always have a corresponding frequency *range*; the term 'frequency' then refers only to the midpoint of that range. *Strictly monochromatic wave trains*[1] *cannot be produced* in a finite experiment.

20.2 The Interference of Wave Groups from Pointlike Wave Centers

Figure 20.1 shows a model experiment (Vol. 1, Sect. 12.12). It illustrates interference between wave groups which are emitted by two centers I and II that oscillate at the same frequency; the groups consist of N "individual waves" or "wavelets", each one with a wave crest and a trough. In this example, $N = 5$. The superposition of these two wave groups gives a simple result: The interference vanishes when the path difference[2] $m\lambda$ of the wave groups becomes larger than the

[1] They should preferably be called *single-frequency*. The term *monochromatic*, that is with a *single color*, is an unhappy choice of wording: *Monochromatic light usually includes a broad range of light frequencies, up to half of the visible spectrum!* (Sect. 29.10).

[2] If one measures Δs, the difference in the lengths of two paths along which two wave groups propagate to the point of observation, as multiples of their wavelength λ, then $m\lambda$ is called their *path difference*.

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_20

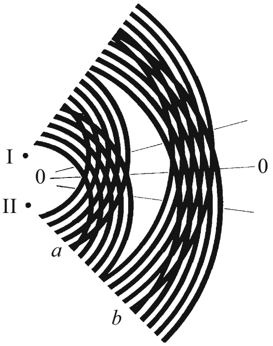


Figure 20.1 The interference of two wave groups: At the left (case a), the groups were emitted simultaneously (phase difference $\Delta\varphi = 0°$); at the right (case b), they are shifted relative to each other by a half wavelength ($\Delta\varphi = 180°$). A mechanical example: At the locations *I* and *II*, individual water droplets fall onto a water surface and produce short groups of capillary waves, like those that everyone has seen when watching raindrops fall onto a puddle

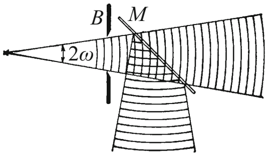


Figure 20.2 Specular reflection by a semitransparent plate *M* (a "beam splitter") can be used to split one wave group into two groups

length $N\lambda$ of the wave groups themselves. Or, put differently, we can observe interference fringes only up to an order $m = N$.

In general, wave groups follow each other without a fixed phase relation. Then the direction of the interference fringes varies randomly between the extremes sketched in Fig. 20.1: Along the line of symmetry $O - O$, for case *a* the wave groups have a phase difference $\Delta\varphi = 0°$; wave crests meet wave crests and troughs meet troughs. In case *b*, the phase difference between the two wave groups is $\Delta\varphi = 180°$, so that wave crests from one group fall on wave troughs from the other. Averaged over time, there are equal numbers of maxima and minima along a given direction, so that we see on the average no maxima and minima (i.e. no interference fringes).

In order to get around this problem, THOMAS YOUNG[3] recognized as early as 1807 that one should not superpose wave groups of the same frequency from *independent* wave centers (emitters), but rather two groups which were emitted as *one* group from the *same* wave center, and were then split and redirected. *To accomplish this redirection,* THOMAS YOUNG *suggested that reflection by mirrors, diffraction,*

[3] THOMAS YOUNG, 1773–1829, studied in Göttingen and lived in London where he had a medical practice. He was a natural scientist with unusually broad interests, and he also made an important contribution to deciphering Egyptian hieroglyphic writing. In 1802, YOUNG was the first to determine the wavelengths of individual spectral regions, by making use of interference fringes in thin wedge-shaped glass plates (Sect. 20.7). He found for example the *wavelengths* at the ends of the visible spectrum to be 0.7 µm (red) and 0.4 µm (violet). He also photographed the interference fringes from ultraviolet light as early as 1803 using paper dipped into a silver nitrate solution! (See R.W. Pohl, *Physikalische Blätter* **5**, 208 (1961)).

and refraction, or some arbitrary combination of these, should all be equally applicable. In Fig. 20.2, for example, a wave group which is incident on a glass plate is "split up" into "transmitted" and "reflected" wave groups. When the reflection occurs at *normal* incidence, the front and the back sections of the same wave group can be superposed to give standing waves. This special case of interference was demonstrated in Vol. 1, in Fig. 12.10 for transverse surface waves on water, and in Fig. 12.47 for longitudinal sound waves in air. A standing electromagnetic wave was demonstrated in this volume in Fig. 12.28 (see also **Video 12.1**). Standing light waves are shown in Fig. 20.25.

20.3 Replacing Pointlike Wave Centers by Extended Centers. The Coherence Condition

Arbitrary phase differences can always be rendered harmless by using YOUNG's method when the wave groups are emitted from a *pointlike* wave center as a statistically distributed sequence over time (Sect. 20.1!). Such a wave center can be either a *single* emitter (Fig. 20.3, top), or many small, neighboring emitters which are independent of each other (Fig. 20.3, bottom). In both cases, there is no difference between the wave groups which propagate along the directions 1, 2, or 3.

This lack of dependence of the wave groups on the direction of propagation of the radiation is lost, however, when the diameter $2y$ of the region over which the numerous emitters are distributed is no longer small compared to the wavelength. Then, an extended wave center of diameter $2y$ can replace a pointlike center only for the radiation emitted within a limited angular range 2ω (Fig. 20.4). Its size is determined by the inequality called the *coherence condition*:

$$2y \sin\omega \ll \lambda/2 \,. \tag{20.1}$$

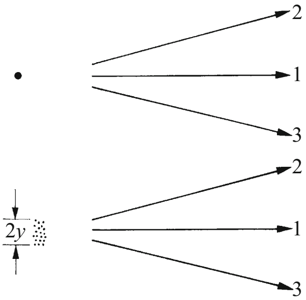


Figure 20.3 Point-like wave centers, i.e. $2y \ll \lambda$. The upper image shows just one center, the lower image shows a number of independent emitters of the same frequency, e.g. the excited atoms in a flame which emit light.

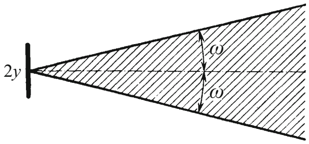


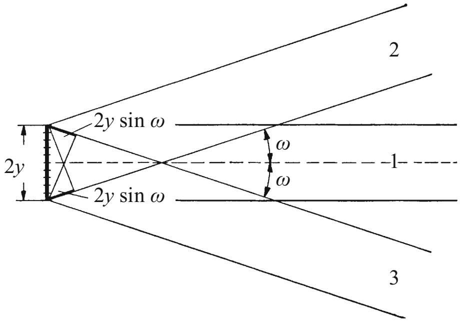
Figure 20.4 The radiation from a light source of diameter $2y$ can be used instead of the radiation from a single pointlike wave center ("point source") only if the opening angle of the light beam obeys the coherence condition $2y \sin\omega \ll \lambda/2$. At this point, we should mention the connection between the coherence condition and the resolving power of a microscope (Eq. (18.5)): One can distinguish an object from its surroundings in a microscope image only if it (when made in some manner to act as a primary light source) projects *incoherent* light through the objective and into the eye of the observer.

It plays an important role, in particular in interference experiments. *If the radiation obeys* Eq. (20.1), *then it is called 'coherent radiation' within this angular range*[4] (**Video 16.3**).

Video 16.3: "Diffraction and coherence" http://tiny.cc/xdggoy The meaning of *spatial coherence* is demonstrated by changing the effective width of the slit which serves as light source (of width $2y$) by rotating it around the optical axis (5:20 minutes).

For the derivation of Eq. (20.1), Fig. 20.5 shows a radiating surface area of width $2y$ (an optical emitter, e.g. a piece of glowing metal, the window of a gas-discharge lamp, or a slit which is irradiated from the left by plane waves from a distant source). We imagine that this emitter surface is divided into individual surface elements marked by dividing lines. Even when there are irregular and unknown variations of the phases and the amplitudes among the surface elements, the resulting radiation will produce plane waves at a distant point along the direction 1, with unknown time variations in their phases and amplitudes. The same is true of an equally distant point along the direction 2. But there, the resulting unknown phases and amplitudes have *different* magnitudes than those at the point in direction 1: The paths traversed by the rays emitted by the different surface elements in direction 2 depend on the locations of the individual surface elements and are not the same as those in direction 1. The path of the ray in Fig. 20.5 emitted by the bottommost surface element along direction 2 is

Figure 20.5 The derivation of the coherence condition



[4] Some authors refer to Eq. (20.1) as the *spatial* coherence condition to distinguish it from a *temporal* coherence condition, $\Delta\nu \cdot \Delta t \ll 1$. This second inequality, however, does not characterize a property of the radiation which is limited to a certain angular range, as in Eq. (20.1). *It simply limits the permissible path difference for the occurrence of interference fringes between the wave groups that are to be superposed.* This path difference must be small compared to the lengths of the wave groups, as shown in Sect. 20.2.

longer by $2y \sin \omega$ than that of the ray emitted by the topmost surface element. This path difference changes the phase φ of the waves in direction 2 by $\Delta\varphi$ as compared to those emitted along direction 1. A phase difference of $\Delta\varphi$ between directions 2 and 1 can be neglected only if $2y \sin \omega \ll \lambda/2$. This is just Eq. (20.1), the coherence condition.

20.4 General Remarks on the Interference of Light Waves

All of what we have discussed in Sects. 20.2 and 20.3 is purely formal geometry; it holds for any kind of waves. With our knowledge of these conditions, we can produce and understand the many forms of interference phenomena with light waves. Although interference phenomena in optics indeed offer nothing fundamentally new, we must nevertheless treat them in detail for three reasons:

1. Interference effects involving light waves play an important role in science and technology.

2. They give rise to long-known phenomena, for example the lively coloration of soap bubbles and thin oil films on water.

3. Using the interference fields of light waves, we can obtain cross-sections on a plane perpendicular to their propagation direction (e.g. a projection screen, a frosted-glass plate, or the object plane of a lens) and can recognize the form of such cross-sections immediately. It is expedient to distinguish among longitudinal, transverse, and oblique observations. These terms are defined in Fig. 20.6.

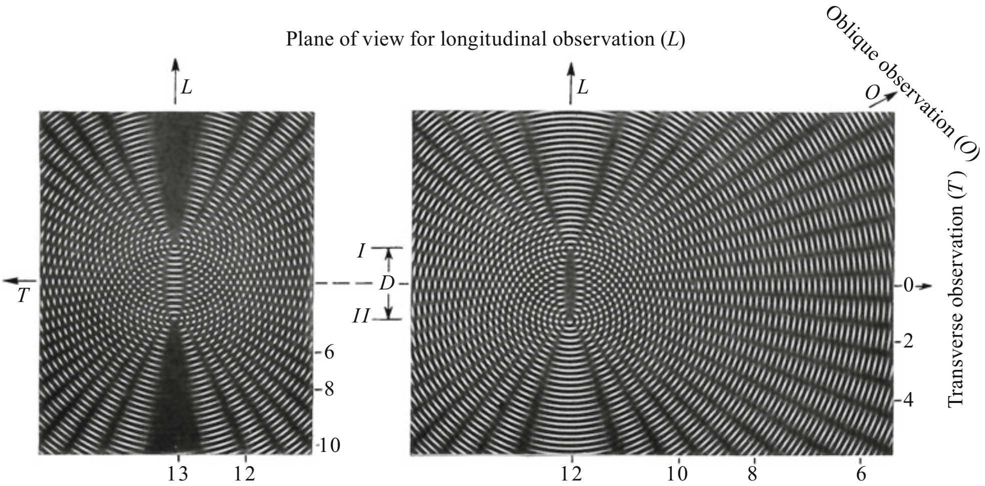

Plane of view for longitudinal observation (L)
L
L
Oblique observation (O)
O
Transverse observation (T)
T
I
D
II
0
2
4
6
8
10
13
12
12
10
8
6

Figure 20.6 Model experiments for the definition of longitudinal (L), transverse (T), and oblique (O) observations of the interference of two wave trains. Two wave trains drawn onto glass plates are projected together. At the right, the spacing D of the two wave centers is an even multiple of $\lambda/2$; at the left, it is an odd multiple of $\lambda/2$. The image at the right was first drawn by THOMAS YOUNG in 1801/02. The numbers on the axes are the *orders*, for minima at the left, and for maxima at the right.

Video 16.3:
“Diffraction and coherence” http://tiny.cc/xdggoy
The video shows THOMAS YOUNG*’s interference demonstration* (beginning at 4.00 min.), using a slit illuminated by a halogen lamp as light source and a CCD camera for registering the image. Both red and blue filter light is used. At the beginning of the video, the experimental setup is explained.

Video 20.1:
“Interference”
http://tiny.cc/9eggoy
The first part of the video also shows THOMAS YOUNG’s interference experiment, but now using a laser as light source and projecting the interference pattern onto a wall for demonstration to a large audience. Interference is observed at the principal maximum of a single slit. (The fact that even with the second slit covered, interference is still seen in one of the secondary maxima, is unavoidable with the simple experimental setup used here.)

C20.1. The concept “spatial” in the characterization of a three-dimensional interference field is in fact superfluous, since fields are always *spatial*. POHL wants to emphasize that the interference occurs in a three-dimensional region and not only on the observation screen.

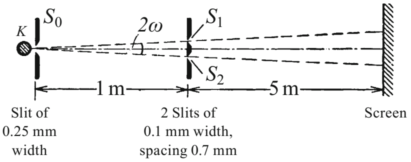


Figure 20.7 THOMAS YOUNG’s interference experiment from 1807 (red filter light, K = arc lamp; see the end of Sect. 20.4. $2y = 0.25$ mm). The interference pattern is shown as a photograph in Fig. 20.9. Here, $\sin\omega \approx 3.5 \cdot 10^{-4}$, so that $2y\sin\omega \approx 10^{-4}$ mm is still smaller than $\lambda/2 \approx 3.5 \cdot 10^{-4}$ mm. (**Videos 16.3 and 20.1**)

To conclude, we give one more *important tip for the experimental demonstration of interference and diffraction phenomena*: One must often make use of irradiated openings (circular apertures or slits) as wave sources instead of pointlike wave centers (Sect. 20.1). Such an opening may however radiate like a point source only within a limited angular range 2ω (which fulfills the coherence condition of Eq. (20.1)!). When the opening is irradiated from all directions, the axes of these angular ranges may be inclined by some angle ω relative to its surface normal (compare Fig. 20.5). Then the surface of the opening perpendicular to the inclined axis ω can be considered to be the pointlike wave center or “point source”.

20.5 A Three-Dimensional Interference Field$^{C20.1}$ with Two Openings as Wave Centers: Transverse Observation

The classical interference experiments, described in 1807 by THOMAS YOUNG[5] (Fig. 20.7), are not only of historical significance, but are also still of practical importance today (Sect. 20.16). YOUNG used two openings (circular holes or slits) to select two wave groups from one original group. These openings, called S_1 and S_2 in Fig. 20.7, serve as wave centers or point sources. They are illuminated from the left by approximately planar waves. These waves are emitted by a light source which is at a distance of around 1 m; it is a lamp that emits light through a slit S_0. One thus obtains two separate beams of light. Using a strictly geometrically-drawn ray construction (the beam axes in Fig. 20.7, dashed), these two beams do not overlap, so that they cannot interfere. In reality, however, both beams diverge as a result of diffraction (Sects. 16.9 and 17.2): Their true distribution is illustrated by the model experiment in Fig. 20.8. Thus, the two light beams overlap in Fig. 20.7 just a few meters beyond the slits S_1 and S_2. From that point on, one can capture the interference fringes on a screen from any point within the spatial

[5] See the footnote at the end of Sect. 20.2, and also Comment C12.3 in Vol. 1.

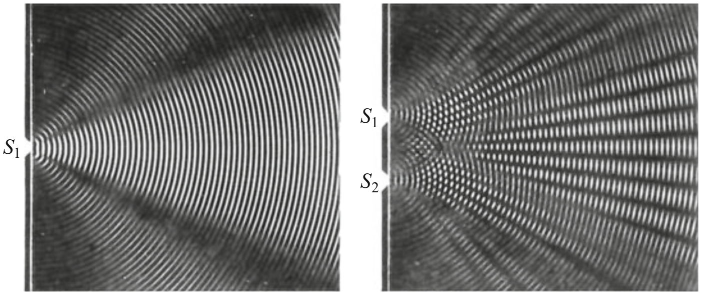


Figure 20.8 Two model experiments illustrating YOUNG's interference demonstration. At the left: The divergent light beam coming from *one* slit; at the right: The superposition of the two beams which emerge from two slits. The left image was obtained in the same way as the one in Vol. 1, Fig. 12.29, right side. In order to produce the image on the right, two glass plates, each containing the left image, were placed one above the other with an appropriate shift.

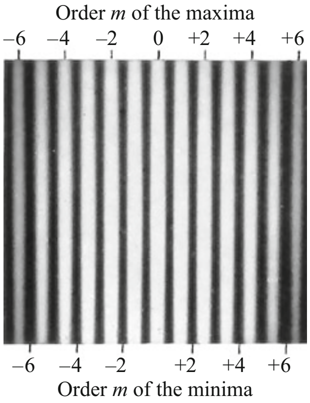
interference field.[C20.1] The fringes shown in Fig. 20.9 (actual size) were photographed at a distance of 5 m in transverse observation. For the angular spacing α_m of the maximum of m-th order, we find:

$$\sin \alpha_\text{m} = \frac{m\lambda}{D} \tag{20.2}$$

(D is the spacing of the two slits S_1 and S_2).

There are numerous variations on this experiment: For example, in Fig. 20.7 we could leave out slit S_0 and replace slit S_2 by a mirror image of slit S_1 (with a mirror in the dot-dashed plane of symmetry in Fig. 20.7. (H. LLOYD, 1837)). Or we could deflect the waves emerging from the slits S_1 and S_2 using flat prisms ("FRESNEL's biprism"), so that they pass along the line of symmetry, simplifying their superposition (A. FRESNEL, ca. 1820). Examples will be described below.

Figure 20.9 A section of the interference pattern (from $m = -6$ to $m = +6$) which is observed with YOUNG's experimental arrangement at a distance of 5 m on a screen at normal incidence (actual size, red filter light, a photographic positive)



20.6 The Spatial Interference Field in Front of a Flat Plate with Two Mirror Images as Wave Centers: Longitudinal Observation

THOMAS YOUNG's interference experiment has a disadvantage: The visibility of the interference fringes is not sufficient to allow it to be demonstrated to a large audience. The diameter $2y$ of the light source must be kept small in order to fulfill the coherence condition. If the width $2y$ of slit S_0 is too large, the fringes disappear. A large-diameter light source requires a very small opening angle ω. This can be obtained using two *mirror images* as wave centers. We produce them in a first experiment by using a flat plate with parallel surfaces.

Figure 20.10 shows a plate (of thickness d) with planar, parallel surfaces at a distance A from an observation screen. K is a lamp which is shielded on its sides and behind by the small box R. The light beam is strongly divergent. It is reflected from both the front and the back surfaces of the plate; therefore, two light beams propagate from the plate to the screen. The two mirror images of the lamp serve as wave centers I and II; circular interference fringes on the screen are the result. The coherence condition need be fulfilled only for *partial beams*, whose opening angle is denoted as 2ω in Fig. 20.10. When the plate is thin, we find

$$\omega \approx \frac{d \sin 2\beta}{2A} . \tag{20.3}$$

Derivation: For sufficiently thin plates and neglecting refraction, we find from Fig. 20.10:

$$\sin 2\omega = \frac{z}{(A + C)/\cos\beta} = \frac{2d \sin\beta \cos\beta}{A + C} = \frac{d \sin 2\beta}{A + C} .$$

For small values of the angle ω, we have $\sin 2\omega \approx 2\omega$, and furthermore, we can neglect C relative to A, thus obtaining Eq. (20.3).

For thin plates, e.g. a mica sheet about 40 µm thick, $\sin\omega$ becomes extremely small, of the order of 10^{-6}. Then the light source can have a diameter of several centimeter and still fulfill the coherence condition (Eq. (20.1)); that is, it acts as a "point source" of light. We could for example use a small Hg discharge lamp; this was the case for the photograph of the interference pattern in Fig. 20.10. It covers the whole wall of a large lecture hall. This impressive demonstration requires no adjustments at all.

Of course, the experiment could also be carried out using a thin *layer of air*. This has the advantage that d can be made still smaller than the thickness

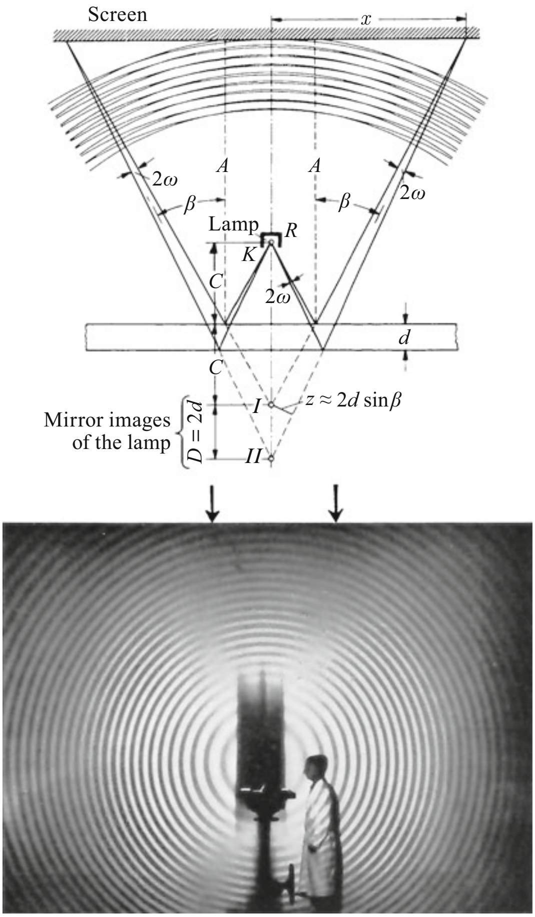


Figure 20.10 The interference experiment illustrated here produces a spatial interference field by making use of a plate with planar, parallel surfaces and divergent light beams. The picture shows a cross-section of the resulting interference field where it intersects the screen (the distance between the lamp and the plate is a few centimeters, while the distance between the lamp and the screen is several meters. Longitudinal observation as illustrated in Fig. 20.6). One can capture "segments" of the interference rings with a matte glass screen close to the surface of the plate if the diameter $2y$ of the lamp is sufficiently small[C20.2] (**Video 20.2, Exercise 20.1**).

> of a mica sheet. Then we can use even a carbon-arc lamp as light source (incandescent light!) Furthermore, with the air layer, the minor disturbance due to double refraction in the mica is absent. (It can be seen in Fig. 20.10 below the arrows at the top of the lower image).

The angular spacing β which belongs to an interference ring of order m, that is the path difference $\Delta = m\lambda$, is denoted by β_{m}. Then for

C20.2. This simple interference experiment, whose pattern covers the whole wall of a large lecture room, is among the most impressive demonstrations of the interference of light waves. POHL published it for the first time in 1940 in *Die Naturwissenschaften* **28**, p. 585. The artful inclusion of his assistant in the picture makes the impressive size of the interference pattern clear even in a book illustration. Later, other textbooks contained references to this experiment, often calling it "POHL's interferometer".
For clarity, the thickness $d = 40\,\mu\text{m}$ of the plate is drawn much too large in comparison to the diameter of the lamp ($\approx$ 1 cm). The mirror images of the lamp are in fact shifted by only a very small distance relative to one another. The diameter $2y$ of the lamp, which enters into the coherence condition, is nevertheless sufficiently small so that one can observe the interference rings quite close to the surface of the plate. The opening angle 2ω of two rays emerging from the lamp becomes greater when the screen or a matte glass disk is brought closer to the plate.

Video 20.2:
"POHL's interference experiment"
http://tiny.cc/kfggoy .

an air layer of thickness d, to a sufficiently good approximation[6] we find:

$$\cos\beta_\mathrm{m} = \frac{m\lambda}{2d}\,. \tag{20.4}$$

The number N of rings is limited. We obtain $N = 2d/\lambda$. The innermost ring has the largest order, namely $m = 2d/\lambda$.

One can observe interference not only with singly-reflected light beams, but also with transmitted beams. Then the direct and the doubly-reflected beams interfere.
The amplitudes of their wave groups are however rather unequal, and the resulting minima are therefore not as dark as with reflected light. This mode of observation can also be used in most of the experiments described in the following sections.

20.7 The Spatial Interference Field in Front of a Wedge Plate with Two Mirror Images as Wave Centers: Oblique Observation

The air layer (or "plate") described above can be conveniently produced with a wedge shape (Fig. 20.11). The light source is at the upper left. The interference pattern in Fig. 20.11 was produced and photographed using this setup. Its area on the wall was around 1 m^2. It is hardly necessary to darken the lecture room for this demonstration.

The air wedge can be replaced by a *soap-bubble film*. It forms a wedge due to gravity, with its thicker base at the lower end.

In contrast to FRESNEL's interference demonstration (see below), the angular extension of the interference field is here independent of ϑ. It is determined by the diameter of the plates. As a result, with the wedge arrangement, we can make the angle $\vartheta \approx \omega$ very small and obtain large, widely visible interference patterns by using light sources with a large diameter $2y$.

C20.3. When a light wave is reflected at the boundary of an optically denser material, i.e. at the boundary of a material with a larger index of refraction, it experiences a phase jump of π or 180°. This phase jump plays a role in several places in the quantitative treatment of interference phenomena involving reflections in the following. The effect is treated in detail in Chap. 25, Sects. 25.7 and 25.8.

Strangely enough, many textbooks still begin with an interference experiment which was described by A. FRESNEL around 10 years *after* TH. YOUNG. FRESNEL's setup (Fig. 20.12) can be obtained from the arrangement sketched in Fig. 20.11 by putting the two reflecting surfaces beside each other rather than behind each other. This arrangement is

[6] This approximation neglects small differences in the angles of inclination β for rays separated only by the angle 2ω, and in addition refraction is neglected, as also in the later Figs. 20.12, 20.13 and 20.32; and finally also the phase jump of the waves upon reflection from an optically denser material, which is quite unimportant in the above connection,[C20.3].

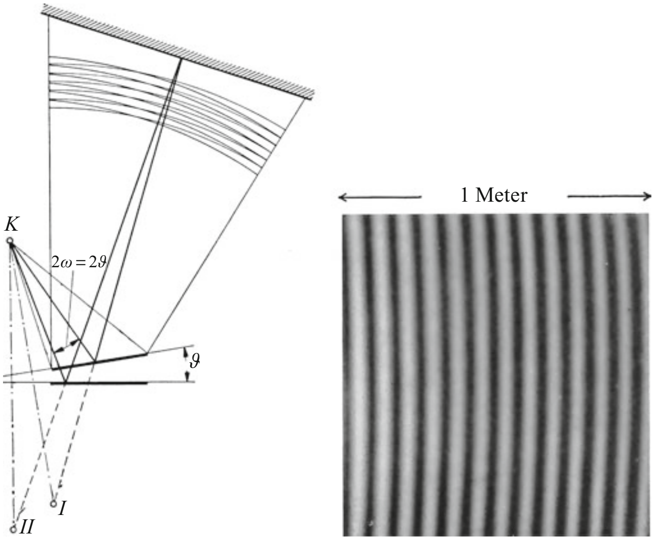
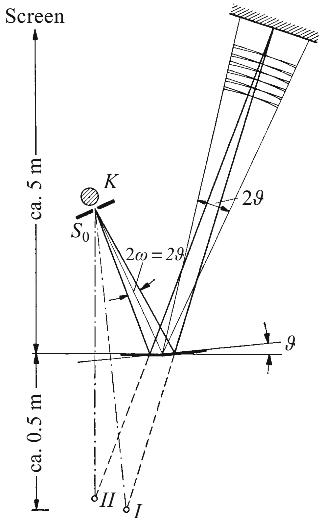


Figure 20.11 An interference experiment with two glass plates placed *one behind the other* (an air wedge). The wave centers I and II are represented here by mirror images of a large light source K (e.g. the crater of an arc lamp with a red filter). *No adjustments whatever need be made.* One places two thick glass plates, for example squares 7 cm on a side (or the bases of two right-angle prisms) on top of each other, with a thin metal-foil strip clamped between them on one side to produce the wedge shape.

Figure 20.12 An interference experiment with two *adjacent* glass plates (FRESNEL's mirror experiment, 1816). The wave centers I and II are represented by the mirror images of a light source in the form of a narrow slit S_0. The *adjustment* is difficult: The surfaces of the two mirrors must not form a step at their point of intersection. This would give an additional path difference and would make the setup usable only with long wave groups, such as those from a sodium-vapor discharge lamp, but not with incandescent light. A section along the length of the image on the screen is similar to the image shown actual size in Fig. 20.9.



however rather disadvantageous for Fresnel's experiment. The angular range of its interference field is only $\approx 2\vartheta$. As a result, the wedge angle ϑ cannot be reduced too far; at the same time, the coherence condition (20.1) must be obeyed for $\omega = \vartheta$. This however sets an upper limit on the diameter $2y$ of the light source S_0, in turn reducing the visibility (intensity) of the interference pattern.

20.8 Interference in the Image Plane of a Pinhole Camera

In Fig. 20.10, the mirror images I and II of the lamp K act as "point sources" of spherical waves. In the figure, on each side two radii of these spherical waves are sketched in as rays (light-beam axes). *The direction of the light in the rays could be reversed.* Then the large screen, illuminated for example by Hg-vapor discharge lamps, would act as the light source. Instead of the "pointlike" lamp K, we then use an accessory which is indispensable for image formation, an *aperture B* (Sect. 17.1). In Fig. 20.13, at the left, we see the opening of a pinhole camera. The aperture B *sorts* the rays which are reflected by the plate (the beam axes) *according to their angles of inclination* β. These in turn determine the path differences produced by the two reflections, which then give rise to interference maxima and minima in the irradiation intensity E_e at the image plane.

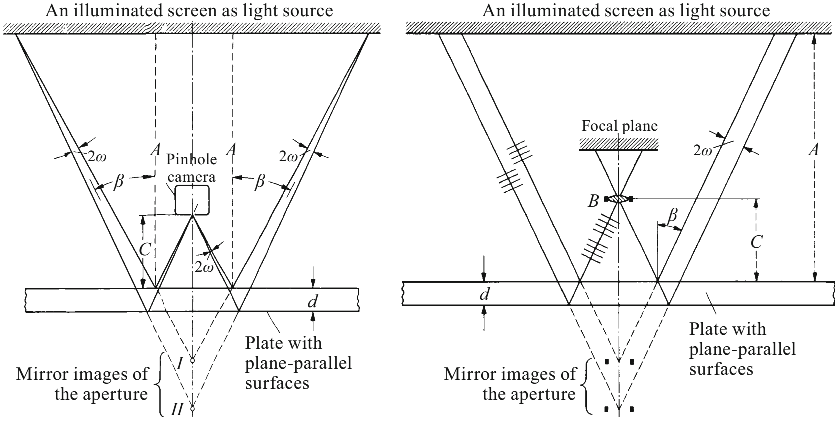


Figure 20.13 The production of circular interference fringes in the image plane behind an aperture B. The shape and position of the extended light source are not important. It could always be replaced by a plane-wave emitter for thought experiments. Here, it is an illuminated screen which is set up parallel to the reflection plate. This setup allows us to clearly recognize the connection with Fig. 20.10.

20.9 Interference in the Focal Plane of a Lens: Longitudinal Observation, Curves of Equal Inclination

Behind the small aperture of a pinhole camera, the luminous density (see Comment C19.7.) of the circular interference fringes is small. Therefore, we replace it by the large aperture of a lens and use its *focal plane* as our *image plane*. In this type of arrangement, the aperture (independently of the thickness of the plate) defines the boundaries of two parallel light beams with the angle $2\omega = 0$. In Fig. 20.13, on the right, besides the axes of the two beams, we see four wavefronts sketched in at two points along the beams.

By far the most convenient arrangement is to use the relaxed lens of an eye.

The walls of the room, its furniture etc. are irradiated with the light from several Hg-vapor discharge lamps and the observer looks at a mica plate (e.g. about 0.15 mm thick) from an arbitrary position. The eye of the observer may approach the plate very closely. *The interference fringes are formed not on or in the plate, but rather on the retina of the observer's eye as the image of an infinitely distant plane. They are completely absent without the observation instrument, in this case the eye of the observer, in contrast to the three-dimensional interference fields in Figs. 20.10 and 20.11.*

An objective observer sees the dark interference fringes as a pattern on a luminous surface, and they seem to be localized on the surface of the reflecting plate. This phenomenon is not explicable in terms of physics, similarly to e.g. the inverted vision in Fig. 15.3.

An analogous example: We see the sky within a small mirror at a distance of ca. 1 m as a bright area. If then for example a wire grid is held midway between the mirror and the eye, we will see the grid as a dark pattern which subdivides the bright area.

To conclude this section, we add a quantitative supplement: The path difference $\varDelta = m\lambda = 2d\cos\beta_\mathrm{m}$ of each pair of rays is given for an air layer by Eq. (20.4). For a layer with an index of refraction $n \neq 1$, we find

$$\varDelta = m\lambda = 2d\sqrt{n^2 - \sin^2\beta_\mathrm{m}} \qquad (20.5)$$

(m = order of the interference maximum = integer. $\varDelta = m\lambda$ holds for maxima, $\varDelta = \left(m - \frac{1}{2}\right)\lambda$ for minima; derivation in Fig. 20.14).

The path difference for a given layer (d = const) is determined only by the angle of inclination β_m. Therefore, we refer to *interference*

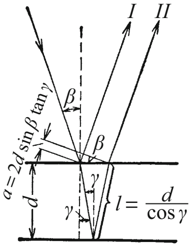


Figure 20.14 The derivation of Eq. (20.5). If the material in the layer or plate has a larger index of refraction n than the material in front of the reflecting surface, then the reflection produces an additional path difference.C20.2 This is only rarely of importance (e.g. in Fig. 21.12) and is thus not taken into account in Eq. (20.5)

curves of constant inclination (W. HAIDINGER, 1849, O. LUMMER, 1884).C20.4 These play an important role in research and technology.

C20.4. See also the LUMMER-GEHRCKE plate in Sect. 22.7.

$$\varDelta = 2nl - a = \frac{2nd}{\cos\gamma} - 2d\sin\beta\tan\gamma\,, \qquad \varDelta = 2d\left(\frac{n - \sin\beta\sin\gamma}{\cos\gamma}\right).$$

Then we set

$$\cos\gamma = \sqrt{1 - \sin^2\gamma} \quad \text{and} \quad \sin\gamma = \frac{\sin\beta}{n}$$

to obtain

$$\varDelta = 2d\frac{n - \dfrac{\sin^2\beta}{n}}{\sqrt{1 - \dfrac{\sin^2\beta}{n^2}}}\,, \qquad \varDelta = 2d\sqrt{n^2 - \sin^2\beta}\,.$$

When the incident and reflected monochromatic light is perpendicular to the plate, that is $\beta \approx 0$, (and is therefore observed only with a lens behind a *small* aperture!), we can consider the interference fringes from layers whose surfaces are not plane-parallel to be *lines of equal layer thickness*. We then refer to *curves of constant thickness*.

With extremely thin layers, e.g. soap-bubble films, oil films on water etc., the interference fringes become very wide. To observe them, we must vary the angle β by a considerable amount according to Eq. (20.5), in order to change the path difference $\varDelta$ by an amount equal to λ and thus pass from one interference fringe to its neighboring fringe. For this reason, when observing with natural light, for example daylight, we often see large surfaces in a single bright color. We then refer to the *colors of thin films*.

The colors of thin films are included in every school physics textbook. Their correct explanation is however more difficult than for any other interference phenomenon. Often, NEWTON*'s rings* are described. Imagine that in Fig. 20.11, the upper surface of the air wedge were replaced by a weakly convex surface which just touches the lower surface at its center. *The explanation becomes simple only for the case that the light falls on the plate or layer at normal incidence, and the plate is viewed perpendicular to its surface.*

20.10 Sharpening the Interference Fringes. Interference Microscopy, MÜLLER's Stripes

In Volume 1, we first discussed the interference of waves from two wave centers (or "point sources"). Then we treated interference of waves from three and more wave centers, and finally from many wave centers arranged on a lattice (Vol. 1, Sect. 12.15). *In this process of increasing the number of wave centers, we found that the interference fringes became sharper and sharper, without changing their positions* (Vol. 1, Fig. 12.40). Experimentally, this was demonstrated using two setups (Vol. 1, Figs. 12.65 and 12.66). For *light waves*, we initially show just one experiment. A detailed discussion will follow in Chap. 22.

Figure 20.15 is an extension of Fig. 20.13. The two surfaces of the plane-parallel plate are made into *semi-transparent mirrors* by adding a layer of evaporated metal. Then, collimated light beams (that is, $2\omega = 0$ from infinitely distant sources) can reach the aperture B after not just two, but a greater number of reflections. Here also, this aperture determines the diameters of the light beams and their angles of incidence β. The essential aspect of this setup is the *lattice-like series of images of the aperture, one behind the other*.

In Fig. 20.15, we see the resulting *dark* fringes on a *light* background. For practical applications, particularly for spectral apparatus, one carries out the observations with *transmitted* light. This yields *light* interference fringes on a *dark* background.

As usual, the interference patterns for incident and transmitted light are *complementary* to each other, i.e. when they are superposed, the fringes cancel each other to give a featureless illuminated area[C20.5] (cf. Sect. 20.12).

How we can carry out the observation using *transmitted light* is described in practical terms in the caption of Fig. 20.15.

Sharp interference curves of this kind are used for example for interference microscopy (J.A. SIRKS, 1893). With this technique, one can investigate samples in which the optical path (Sect. 16.3), that is the product of the index of refraction n and the layer thickness d, varies somewhat from one microscopic region to another. A simple example (for the special case of $n =$ const) is the measurement of the thickness of thin films (Fig. 20.16). A shift of an interference fringe by $1/100$ of the fringe spacing corresponds to a height difference of $3 \cdot 10^{-9}$ m $= 3$ nm (using *monochromatic* illumination, $\lambda \approx 600$ nm).

For the application of interference microscopy, an old technique has again become interesting, namely the spectral decomposition of interference fringes which were produced using *incandescent light* and lie transverse to the long axis of the entrance slit of the

C20.5. To explain this complementarity, remember that looking from above, one of two interfering light waves experiences a phase shift of 180° (at the boundary of the optically denser material), while for the transmitted light, there is no phase shift (see Comment C20.2). How such a phase shift can exchange light and dark within an interference pattern can be very clearly seen in YOUNG's two-slit experiment (Fig. 20.8, right): Imagine that the waves emerging from one of the slits are delayed by a phase of 180°; then the light and dark regions are exchanged.

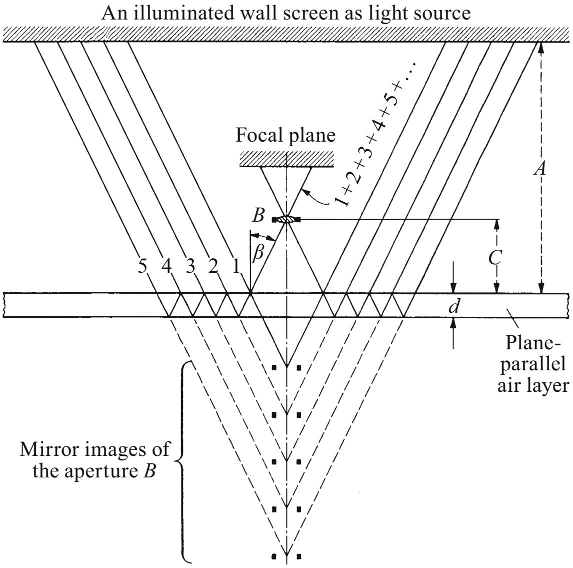

An illuminated wall screen as light source
Focal plane
1+2+3+4+5+...
B
β
A
C
d
5 4 3 2 1
Plane-parallel air layer
Mirror images of the aperture B

Figure 20.15 The sharpening of the interference fringes through the use of many collimated light beams with equal path differences between neighboring beams (for an air layer, this difference is $2d \cos \beta$). The sketch applies to the observation of concentric circles (*curves of constant inclination*) with *reflected* light. If *transmitted* light is used, one exchanges the lens for the mirror image of the aperture closest to the plate. In this way, for the case of a plate in the form of an air layer several centimeters thick, we arrive at the scheme of the *high-resolution spectral apparatus* which was described in 1897 by CH. PÉROT and A. FABRY. The thick air layer is located between the two mutually parallel, *semi-transparent* mirror surfaces of two glass plates (which are slightly wedge-shaped to avoid disturbing reflections). The illuminated screen is usually replaced by a condenser lens with the light source that is under investigation in its focal plane. More details can be found in Sect. 22.7.

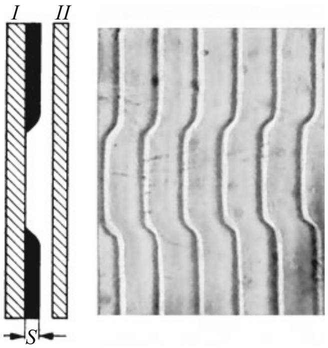


Figure 20.16 Interference microscopy. At left we see the stepped profile of a thin air layer which produced the interference pattern seen at the right. The thickness of the evaporated layer is $S = 0.1\,\mu\text{m}$. For clarity, the silvering on the plates is not drawn. As in Fig. 20.13, these are curves of constant inclination, with a large angle of inclination β, observed with reflected light. The dark interference fringes on a light background are segments of large circles (use a ruler to check this!). (**Exercise 20.2**)

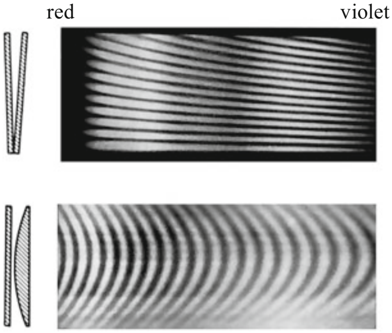


Figure 20.17 On the right: MÜLLER's stripes in a continuous spectrum; left: The profiles of the thin air layers which produced the interference patterns. (First published in 1807 by Th. YOUNG as hand-colored drawings)

spectral apparatus. In this method, the continuous spectrum of these "MÜLLER's stripes" is scanned. These are *colored curves of the same fringe order m*. Each profile transverse to the slit axis corresponds to a particular shape of MÜLLER's stripes. Figure 20.17 shows two examples. Stepwise changes in the layer profile produce steps in the stripes. When the surfaces are silvered, MÜLLER's stripes also become extremely sharp. Using these stripes, S. TOLANSKY (1907–1973) was able to measure steps on single-crystal surfaces of 1 nm, i.e. of molecular dimensions, with an incandescent light source. He achieved the power of electron microscopy by using simple interference microscopy.[C20.6]

C20.6. For measurements of layer thickness in the nanometer range, in addition to interference methods, various other techniques are used, among others the quartz-oscillator balance method and scanning tunnelling microscopy (STM). Each of these techniques has its specific advantages and disadvantages, depending on the application intended.

20.11 The Lengths of Wave Groups

Using red-filter light, we can observe interference fringes out to an order of $m = 10$, i.e. with a path difference of $\varDelta = 10\lambda$. From Sect. 20.2, this allows us to draw conclusions about the lengths of the wave groups: The wave groups of red-filter light must consist of about $N = 10$ individual waves (i.e. wavelets, each with "one crest and one trough").

Interference fringes of much higher orders m, with path differences $\varDelta$ of up to many thousands, sometimes even more than $10^6\lambda$, can be obtained with the radiation emitted by some metal vapors excited electrically or thermally to light emission. This can be rather conveniently seen using the light from technical Na-vapor lamps (an electric arc between electrodes made not of carbon, but rather of sodium). We can attribute wave groups of considerably greater length to such light sources; in the visible, they are in the range of 0.1 mm up to 1 m. They consist of around $1.5 \cdot 10^2$ to $1.5 \cdot 10^6$ individual wavelets. Light with long wave groups is called "monochromatic". *Wave trains of practically unlimited length are emitted by the light sources called* "lasers".[C16.4]

How should we imagine the wave groups of incandescent light, that is the light from a glowing solid body (an arc lamp, an incandes-

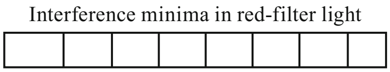


Figure 20.18 Interference fringes, produced using red-filter light with an air wedge 28 mm long, whose thickness increases up to 10^{-3} mm. Both plates are rectangular and are ground to be flat wedges, in order to avoid disturbing reflections. In addition, a linear facette is ground into the upper plate so that the plates remain in "optical contact" at the thin edge of the wedge. The width of the minima is reduced by multiple reflections.

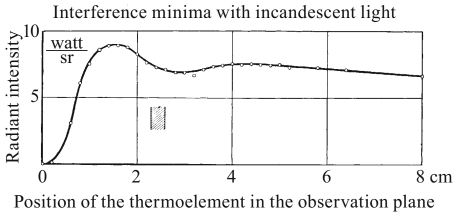


Figure 20.19 The same interference pattern as in Fig. 20.18, observed using incandescent light and a thermoelement (as a non-selective receiver or radiometer) in the plane of the observation screen. Its width is indicated by shading. At 0, the radiation passes without reflection through the plates, which are in "optical contact" at this point. (sr means *steradian*, the unit of a solid angle; see the footnote in Sect. 19.2 and Vol. 1, Sect. 1.5)

cent filament lamp, the tiny carbon particles in the hot flame gases of a candle, etc.)? To answer this question, the simple interference setup described in in Fig. 20.11 can be used; however, the plates are now made not of glass, but rather of lithium fluoride. This material is superficially similar to glass, but it transmits not only visible radiation, but also the neighboring spectral regions, the infrared and the ultraviolet[7]. At the sharp edge of the air wedge, the two plates are in "optical contact"; there, no reflection occurs and there is no splitting of the light into two partial beams.

To test the setup, we first employ red-filter light. Dark interference fringes on a light background appear on the screen (Fig. 20.18).

For the observations, instead of the eye of an observer, we use a thermoelement which has been blackened with soot (Fig. 15.5), and is thus a physical radiometer[8]. The results of the measurements are shown graphically in Fig. 20.19. The zero point of the abscissa marks the position of the "optical contact". Starting there, the radiant intensity I_ϑ (power/solid angle) of the radiation reflected from the two wedge surfaces increases until it reaches an approximately constant value: Instead of many interference fringes, we find in Fig. 20.19 only two flat maxima. The result: An interference field produced

[7] Compare Fig. 27.1 for NaCl, the best-known of the alkali halides.

[8] This radiometer is therefore not selective for radiations in different wavelength regions like the human eye, which reacts to some regions in varying ways (chromatic hues! Sect. 29.9), and to some regions not at all.

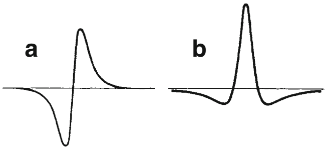


Figure 20.20 Examples of two short, nearly aperiodic wave groups. An unstructured series of such groups could be used to describe the behavior of incandescent light in interference experiments.

with incandescent light shows only a weakly-developed structure. Incandescent light behaves as though it could be described as an unstructured series of short, nearly aperiodic wave groups, as shown for example in Fig. 20.20. The true wave picture of incandescent light can be imagined as being similar to a noise spectrum in acoustics. More details will be given in Sect. 22.4.

20.12 Redirection of the Radiant Power by Interference

Interference cannot create or destroy the radiant power (or energy current), but instead just redirects it. What is lacking in the direction of the interference minima is added in the directions of the interference maxima. We offer two technically important examples of this:

1. *Eliminating reflections, non-reflective coatings.* A single glass surface reflects about 4 % of the normally-incident radiant power, while about 96 % passes through the surface and into the glass. Imagine a thin, evaporated crystalline film on the glass block G in Fig. 20.21. Its material is chosen so that the indices of refraction $n_{\text{air-film}}$ and $n_{\text{film-glass}}$ are approximately equal in some spectral region. Then the same fraction of normally-incident radiant power is reflected at the boundaries 1 and 2. In addition, the thickness d of the crystalline film is adjusted so that the two wave trains which are reflected upwards have a path difference of $\Delta = \lambda_m/2$ for an average wavelength of λ_m within that spectral region. Then they will cancel each other through interference. The surface is thus strictly non-reflecting for λ_m, and all the radiant power which falls at normal incidence on the boundaries 1 and 2 passes without reflection losses into the glass block G. For the neighboring spectral regions on each side of λ_m, the thickness d is only approximately equal to $\lambda/2$, so that the suppression of reflection is incomplete, but still, for many practical purposes, it is sufficient (compare Sect. 25.8).
2. *Layered mirrors with nearly loss-free reflection; reflection filters.* Instead of providing a non-reflective coating, we can redirect the large fraction of light that penetrates into the glass block in spite of reflection losses,

Figure 20.21 Non-reflective coating

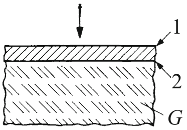


Figure 20.22 The structure of a reflection filter

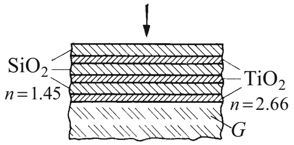
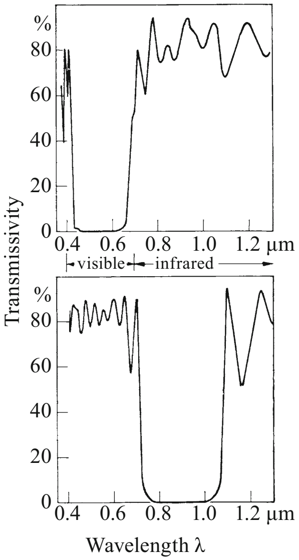


Figure 20.23 Top: A reflection filter which allows only small portions of the visible spectral region to pass, in the violet and the red. Bottom: A reflection filter which passes around 80 % of the light in the visible, but no infrared between $\lambda = 0.8\,\mu$m and $\lambda = 1\,\mu$m.



that is roughly 96 % of the radiant power can also be reflected. To achieve this, we need a large number of reflecting ancillary surfaces. They can be obtained by alternately evaporating two types of thin crystalline films (Fig. 20.22). Each boundary surface, as an ancillary mirror surface, reflects the same fraction of the normally-incident radiant power. The layer thicknesses are chosen so that the path difference $\varDelta$ for the average wavelength λ_m of the spectral region that is to be reflected is equal to λ_m[9] With $\varDelta = \lambda_\mathrm{m}$, the amplitudes of the reflected wave trains all add with the same phase. In this manner, we can fabricate nearly loss-free mirrors for selected, narrow spectral regions; in the visible region, these are vastly superior to metallic mirrors. For broad spectral regions, by using around 20 to 30 layers we can obtain *reflection filters* which remove certain spectral regions not through absorption, bur rather by reflection. There are for example reflection filters which allow no visible light to pass through (Fig. 20.23, top); or no infrared light from the spectral region adjacent to the visible (Fig. 20.23, bottom).

20.13 Interference Filters

Like all spectral apparatus, plane-parallel plates (FABRY-PÉROT étalons) can also be used to separate narrow spectral regions out of incandescent or natural light. In this way, we arrive at “interference

[9] Jumps of $\lambda/2$ are included in the path differences $\varDelta$; these are due to reflection by a more optically dense material.$^{\mathrm{C20.2}}$

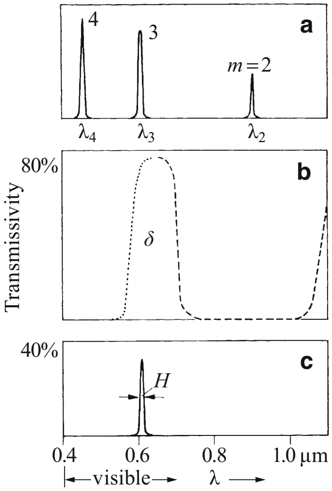


Figure 20.24 An interference filter. A shows the light from the interference plate alone; in B, an absorption filter (e.g. colored glass) for short wavelengths *dotted*, and a reflection filter for long wavelengths *dashed*. In part C, we see the combination of A and B (the half-width of the transmitted spectral region around $\lambda = 600\,\text{nm}$ is $H = 10\,\text{nm}$, so that the spectral selectivity λ/H is equal to 60).

filters" (usable only for light at normal incidence). Their principle: In Fig. 20.15, we replace the plane-parallel air layer by a very thin ($d < 10^{-3}$ mm), non-absorbing crystalline film, e.g. a film of MgF_2, partially silvered on both surfaces. For incandescent light at *normal incidence* ($\beta = 0°$), it permits only narrow spectral regions to pass through, whose wavelengths λ_1:λ_2:λ_3:... have the ratios 1:$\frac{1}{2}$:$\frac{1}{3}$:... They emerge as bright interference maxima on a dark background, with the orders $m = 1, 2, 3, \ldots$ (Fig. 20.24 A). These plane-parallel crystalline films can be combined with suitable filter layers so that only one of the transmission regions remains, e.g. at λ_3, as in Fig. 20.24 C.

20.14 Standing Light Waves

(Otto Heinrich Wiener,[C20.7] 1890). The wavelengths of visible light are only some few 10^{-4} mm. Nevertheless, we can obtain *standing light waves* – although not readily in simple demonstration experiments – by using the technique described in Vol. 1 (Sect. 12.5).

For example, one can press a liquid mercury mirror against an extremely fine-grained photographic plate. The light which arrives at this mirror at normal incidence and is then reflected by the mercury darkens the photographic emulsion in equidistant, separated layers, with a spacing of $\lambda/2$. Figure 20.25 shows a thin slice from such an emulsion, cut perpendicular to the plane of the plate, at a high magnification.

[C20.7] See H. Jäger, *Annalen der Physik*, 5th series, Vol. 34 (1939), p. 280.

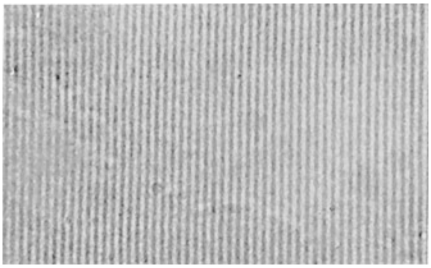


Figure 20.25 The photographic detection of standing light waves ($\lambda = 546\,\text{nm}$). A magnified view of a section through a gelatine emulsion, allowed to swell up to 10 times its original thickness by immersing in water. This is a small section from an image produced by S. MAGUN in 1935; the image makes several hundred wave crests and nodes visible (cf. Vol 1, Fig. 12.47).

20.15 Interference due to Particles which Redirect the Light

In Figs. 20.10 and 20.13, each ray coming from the light source was redirected at the two surfaces of a plane-parallel plate by *reflection* and thus split into two partial rays. Instead of reflection, the rays can also be redirected by *diffraction* or *scattering*. A redirection by small diffracting or scattering particles can for example also be used, when these particles are arranged directly in front of or on the surface of the plate.

One could, for example, take an ordinary household mirror (i.e. a glass plate, by no means plane-parallel, which is silvered on its rear face) of around 30 cm diameter. The glass surface is dusted or rubbed with modelling clay. A small light source is set up about 2 m in front of the mirror, and the eye of the observer is at some arbitrary distance behind the light source. In Fig. 20.26, we have allowed ourselves a certain luxury: The arc lamp is placed at the side and sends its light via a small metal mirror *H* onto the dusted surface of the large mirror. This allows us to place our eyes practically "at the position of the light source". Looking at the mirror in a direction *perpendicular* to it, we can see concentric circular interference rings on its surface. They are surprisingly clear. Behind their common center, we see the image of the light source. The diameter of the rings varies with the distance of the observer from the mirror. In red-filter

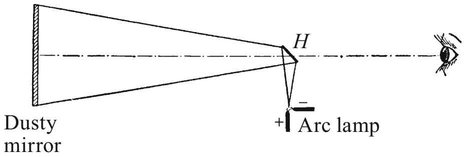


Figure 20.26 Subjective observation of the interference rings which are produced on the surface of a thick household mirror after dusting or clouding it. They are often called "QUETELET's rings" (but they were described already in 1704 in NEWTON*'s "Opticks"* in great detail). *H* is a small planar mirror which acts as light source.

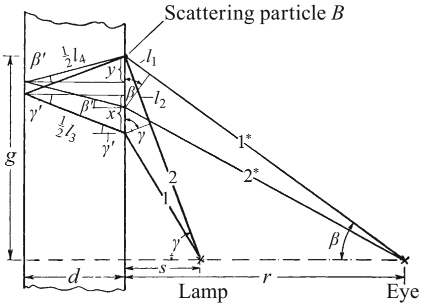


Figure 20.27 The occurrence of small path differences in thick mirror-glass plates (refraction was neglected here), and the derivation of Eq. (20.6). The decisive angle for the coherence condition, 2ω (denoted here as γ), is quite small. For clarity, the angles β and γ are exaggerated in the drawing. However, for the calculation, we assume (as in the actual experiment) that the angles are small. We thus set $\sin\beta = \tan\beta$, etc.

light, we can easily count the usual 10–15 orders. When we view the mirror obliquely, the center of the rings appears shifted. In incandescent light, we see a bright, colorless ring of *zeroth* order; behind it is the image of the lamp. The rings bordering this central maximum appear deep black to the eye. They are followed by the usual series of higher-order rings, in gradually fading colors.

Interference fringes of low order can be produced only by *small* path differences. How can they occur here in spite of the thick mirror-glass plate? Answer: As small *differences* between two large path differences (Thomas Young, 1802). In Fig. 20.27, B is one of the large number of particles which redirect the light rays reaching the mirror. Two paths lead both from the light source to the particle B and also from the particle to the eye of the observer. Along path 1, the light from the source arrives at the particle B via a 'detour'. From B, however, after being redirected by diffraction or scattering, it reaches the eye directly along the path 1^*. Along path 2, the light from the source arrives directly at B, but from B, after redirection by diffraction or scattering, it again takes a 'detour' along path 2^* to arrive at the eye of the observer. The path difference $\varDelta$ between these two wave trains is thus rather small. We define the ratio

$$\frac{\text{Distance } r \text{ to the eye}}{\text{Distance } s \text{ to the lamp}} = q\,.$$

Then, for small values of β and γ, and in the limiting cases $q \gg 1$ or $q \ll 1$, we have

$$\varDelta = \frac{d}{n}\left(q^2 - 1\right)\sin^2\beta \tag{20.6}$$

(Derivation see below; nomenclature as in Fig. 20.27).

At the m-th maximum, $\varDelta = \pm m\lambda$; then for its angular spacing, we have

$$\sin^2\beta = \pm\frac{m\lambda \cdot n}{d\,(q^2 - 1)} \tag{20.7}$$

(β is the angle of inclination as in Fig. 20.27, d the thickness, and n the index of refraction of the mirror-glass plate. The minus sign holds for

$q < 1$). The same angle β therefore appears for two different values of q. In one case, the eye is in front of the light source, in the other case, it is behind the source (as in Fig. 20.26). The innermost ring has the smallest order m, while in Fig. 20.10, it had the largest.

The difference in the optical paths (Sect. 16.3) determines the path difference between the two wave trains:

$$\Delta = (l_2 + nl_4) - (l_1 + nl_3), \tag{20.8}$$

$$l_2 = 2x \sin\gamma, \quad x = d\tan\gamma' = d\sin\gamma' = \frac{d}{n}\sin\gamma,$$

$$l_2 = \frac{2d}{n}\sin^2\gamma \quad \text{and analogously,} \quad l_1 = \frac{2d}{n}\sin^2\beta,$$

$$nl_3 = \frac{2dn}{\cos\gamma'} = \frac{2dn}{\sqrt{1-\sin^2\gamma'}} = 2dn\left(1 + \frac{1}{2}\sin^2\gamma'\right),$$

$$nl_3 = 2dn\left(1 + \frac{1}{2}\frac{\sin^2\gamma}{n^2}\right) \quad \text{and} \quad nl_4 = 2dn\left(1 + \frac{1}{2}\frac{\sin^2\beta}{n^2}\right),$$

$$\Delta = \frac{d}{n}(\sin^2\gamma - \sin^2\beta).$$

For small values of the angles β and γ, we find

$$\frac{\sin\gamma}{\sin\beta} = \frac{\tan\gamma}{\tan\beta} = \frac{r}{s} = q, \quad \Delta = \frac{d}{n}(q^2 - 1)\sin^2\beta\,. \tag{20.9}$$

20.16 YOUNG's Interference Experiment in the FRAUNHOFER Limit

In YOUNG's interference experiment (Fig. 20.7), the wave centers are two openings (holes, or better, slits), S_1 and S_2. Their spacing D can be chosen to be at most a few millimeter in the simplest setups; otherwise, the two light beams will no longer overlap. This small slit spacing is often annoying. However, we can free ourselves from this limitation and use any values of the slit spacing D that we care to: We need only set a lens L_1 behind the slits S_1 and S_2. This is illustrated in Fig. 20.28. The lens L_1 redirects the two light beams which emerge from the slits S_1 and S_2 as divergent beams; they now converge towards the optical axis. They then intersect at the image

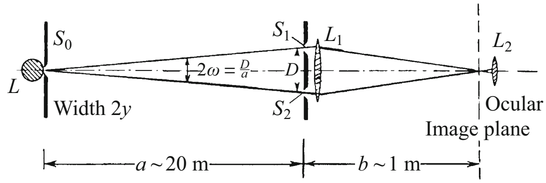


Figure 20.28 YOUNG's interference setup with the FRAUNHOFER observation condition (L is a metal-vapor discharge lamp (Na or Hg)). The lengths of the distances a and b permit readily-measurable widths $2y$ to be used for the slit S_0 (**Videos 16.3 and 20.1**).

In **Video 16.3**, "**Diffraction and coherence**" http://tiny.cc/xdggoy, after 4:00 min.; and in **Video 20.1, "Interference"** tiny.cc/9eggoy, which shows YOUNG's interference experiment in its first part, however with FRESNEL's observation condition (see also Sect. 20.5).

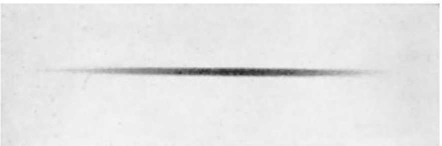


Figure 20.29 A single-slit diffraction pattern, observed using YOUNG's setup

plane with practically planar wavefronts, but the wavefronts are more strongly tilted relative to each other than without the lens. Therefore, the resulting interference fringes are more closely spaced than they were without the lens. We can observe the fringes either through a magnifying lens L_2 (ocular or TV camera), or else project them as an enlarged image onto a frosted glass screen using an objective lens. The lens L_1 and the magnifying glass L_2 (ocular) together form a *telescope*. Indeed, one usually makes use of a telescope with two slits S_1 and S_2 in front of its objective. In this form, YOUNG's experimental arrangement is especially important. We will thus treat it in some detail.

Initially, we set the slit S_0 to a narrow width and cover either slit S_1 or slit S_2. In both cases, we obtain the *diffraction pattern* as photographed in Fig. 20.29, and it is at the same position in the image plane, symmetric to the optical axis. It is the central maximum of the diffraction pattern that we have seen in Figs. 17.6 and 17.7 (its secondary maxima are too faint to see).

Next, both of the slits S_1 and S_2 are opened at the same time. The wave trains coming from S_1 and from S_2 interfere with each other: The diffraction pattern is sliced through by sharp interference fringes (Fig. 20.30). Here, an important precondition was met: The slit S_0 acted in spite of its finite width $2y$ like a "point" (or rather a line-shaped) light source. The slit width $2y$ thus meets the coherence condition

$$2y \sin\omega \ll \lambda/2\,. \tag{20.1}$$

Now we describe something new: the

Measurement of the diameter of a distant light source

(A.H.L. FIZEAU, 1868).[10]

We gradually increase the width of slit S_0 and thereby violate the coherence condition. Nevertheless, we still see the interference fringes, however more faintly, i.e. with poor contrast between a maximum and its neighboring minimum. This decrease in the contrast is easy to understand: Imagine that the slit S_0 were sliced up into small segments along its length. Each one gives rise to an interference pattern as in Fig. 20.30, but the individual interference patterns are shifted along the axis of the slit relative to each other. Their superposition gives rise to washed-out fringes.

[10] *Comptes rendus, Paris*, **66**, 934 (1868), and J.M. Stephan, *ibid.*, **78**, 1008 (1873).

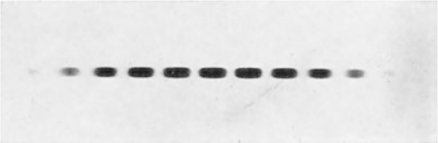


Figure 20.30 The interference fringes observed within the diffraction pattern from Fig. 20.29: Single-slit diffraction and two-slit interference are superposed

When

$$2y \sin \omega = \lambda \,, \tag{20.10}$$

the interference fringes have completely disappeared, *but only temporarily*: As the slit width $2y$ is increased still further, they return, even more washed-out or lacking in contrast than before. For $\sin \omega = 2\lambda/2y$, they again disappear, and so on through several repetitions (that is for $\sin \omega = 3\lambda/2y$, etc.). This is termed *partial coherence*.

From Eq. (20.10), setting $\sin \omega = D/2a$, we find

$$2y = \frac{2a\lambda}{D} \quad \text{or} \quad \frac{2\lambda}{D} = \frac{2y}{a} = 2\varphi \,, \tag{20.11}$$

where 2φ is the angle of vision subtended by an object of diameter $2y$ seen from the distance a. With this relation, we can refer the unknown diameter $2y$ or angle 2φ of a distant light source to known quantities, and thus measure it.

A.A. Michelson made use of Fizeau's technique to determine the angles of vision 2φ of several nearby fixed stars at known distances a from the earth, for example that of α Tauri (Aldebaran), for which he found $2\varphi = 0.020$ seconds of arc. To achieve this, the distance D between the two slits S_1 and S_2 in Fig. 20.28 was varied in a measurable way. Using mirrors, D could be made even greater than the diameter of the telescope objective.

20.17 Optical Interferometers

Optical interferometers are used to accomplish two tasks:

1. For highly precise comparisons of lengths or distances (e.g. length measurement standards) on the one hand, and the wavelength of the light on the other (cf. Vol. 1, Sect. 1.3).

2. For comparisons of two coherent light beams with different histories, e.g. after they have passed through different materials.

The simplest but already quite serviceable interferometers make use of transverse observation. They employ the basic experimental setup

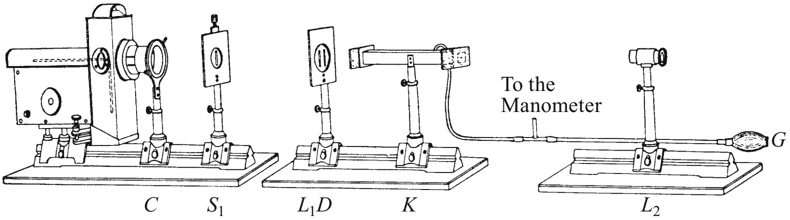
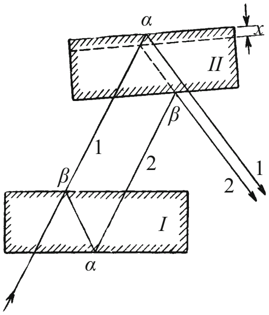


Figure 20.31 An interferometer for determining the index of refraction of gases at various densities (schematic in Fig. 20.28; G is a rubber bulb for changing the gas density ϱ). The two slits D immediately behind the lens L_1 ($f = 2$ m) are 2 mm wide. Their spacing is 10 mm. The distances S_1L_1 and L_1L_2 are about 4 m. Both light beams pass through the glass window K which closes off the end of the gas container and is sufficiently wide to intercept the collimated reference beam in the air beside the gas vessel at K.

Figure 20.32 An interferometer with two collimated light beams which are shifted sideways and parallel to each other. The effective thickness x can be varied by tipping the plates relative to each other. When they are parallel, x and thus also the path difference between the two beams 1 and 2 are equal to zero. All unnecessary reflections, which in reality are eliminated by apertures, have been left out of the drawing.



of Thomas Young in the arrangement shown in Fig. 20.28. There, the two light beams are separated transversally by several centimeter just behind the lens. It is thus easy to allow one beam to pass through the air and the other through some other gas, in order to compare the wavelengths in the two gases. Experiments of this kind are discussed below; Fig. 20.31 shows a practical example.

All of the other interferometer designs also make use of interference fringes in the image plane of a lens (often the lens of the eye), but they employ longitudinal observation (Fig. 20.6). They implement a *plane-parallel* plate of thickness x as the *difference* between two plates of unequal thicknesses (Th. Young, 1817). This is illustrated for example in Fig. 20.32, where the axes of two light beams which are shifted parallel to each other are drawn. Often, one replaces the plates completely or partially by mirrors (e.g. at α), or partially-silvered mirrors (at β). One thus arrives at the interferometer design of Albert A. Michelson (Fig. 20.33), with two mutually perpendicular light beams (cf. Vol. 1, Fig. 12.67). The path difference is denoted by x. By tipping the mirror II, one can also produce a *wedge* plate. The plate III is in fact not necessary in principle, but it allows one to achieve equal path lengths through glass for both light beams. That simplifies the observations.

C20.8. Today, similar interferometers with arm lengths of up to several kilometer are in use, for example in the gravitational-wave detector system LIGO-Virgo. (See e.g. https://en.wikipedia.org/wiki/LIGO). Still much longer arm lengths are planned for the space version of the experiment, "Cosmic Explorer".

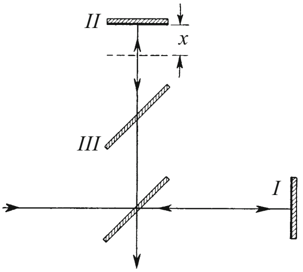


Figure 20.33 The interferometer of MICHELSON (only the axes of the light beams are drawn). In the largest examples of this design, the mutually perpendicular light paths ("arm lengths") are 30 m long.[C20.8]

20.18 Coherence and Fluctuations in the Wave Field

In Fig. 20.34, A represents an aperture which radiates in all directions and has a diameter of $2y$. The whole solid angle from its center to the screen S can be subdivided into small solid-angle elements, within each of which the coherence condition (Eq. (20.1)) is fulfilled. Some of them, chosen randomly, are marked in the figure as 1, 2, … The radiation which propagates within these solid-angle elements strikes the screen in the regions $I, II, \ldots$ The phase distribution in the emitting aperture A can initially be thought of as arbitrary, but constant over time. Fulfilling the coherence condition means that the aperture acts as a *point wave source* within each of the individual angular elements 1, 2, … As a result, sections of the regions $I, II, \ldots$ cannot be irradiated with different intensities; rather, the irradiation of each individual surface element $I, II, \ldots$ must be *uniform*. Phase changes of the individual emitting elements within A can change the radiant intensity in each of the surface elements $I, II, \ldots$ only in a *unified* manner. Such changes may be of different strengths for the different surface elements $I, II, \ldots$. Within each of these elements, the radiant intensity can vary between zero and a maximum value. Statistical phase changes within the emitting area therefore produce *fluctuations* on the screen S.

In the wave field of the light, these fluctuations occur much too quickly to be observable with simple techniques. But they can be simulated quite successfully in model experiments, in the simplest case using subjective observation: We look through a red filter and a *moving* piece of frosted glass at a single, distant light source (W. Martienssen and E. Spiller, *American Journal of Physics* **32**, 919 (1964))[11]

Suppose that the emitting surface A is a sheet of white paper which is illuminated with the extremely monochromatic light beam from a laser. Then *transverse* variations of the phase distribution are absent, and with them the fluctuations: Their spots on the screen are "frozen in"; instead of a fluctuation, the screen shows a *granulation*. This must be taken into account in making photographic images (e.g. for holography).

Our understanding of fluctuations and granulation can be increased by considering Fig. 12.49 in Vol. 1. There, a hand at rest in the wave field produces a granulation, while a hand which is changing its shape randomly produces fluctuations. Both are made visible in the figure by using the artifice of the acoustic replica method (Vol. 1, Sect. 12.18).

[11] To show this demonstration to a large audience, the surface of a glass plate is covered with a layer of glued-on glass powder and then moved back and forth perpendicular to the beam axis of a beam of light which is projected onto a wall screen.

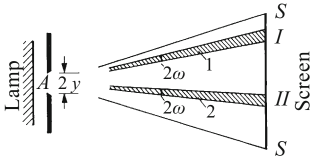


Figure 20.34 The origin of fluctuations in the wave field

Exercises

20.1 The interference experiment shown in Fig. 20.10 is carried out with an air layer of thickness $d = 40\,\mu\text{m}$. The wavelength of the light used is $\lambda = 600\,\text{nm}$. a) What is the value of the largest order m_{max}, and how large is the radius x of the corresponding interference ring, when the distance of the plate from the wall of the lecture room is $A = 3\,\text{m}$?
b) Given this distance, how large is the radius of the tenth interference ring x_{10} (counting from inside to outside)? (Sect. 20.6)

20.2 The application of interference microscopy to the determination of the thickness S of an evaporated metal film gives a shift in the interference fringes of 1/3 of the spacing between two fringes whose orders m differ by 1 (Fig. 20.16). The wavelength of the light used is $\lambda = 600\,\text{nm}$. How can we calculate the film thickness S from this, and what is its value? (Sect. 20.10)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_20) contains supplementary material, which is available to authorized users.

Diffraction 21

21.1 Casting Shadows

The diffraction of light as an extension and softening of the geometrical shadow boundaries has been treated already in Sects. 16.9 and 17.3. In this chapter, we will investigate the phenomenon of diffraction in some more detail.

Diffraction of mechanical waves was treated in depth in Vol. 1 (Chap. 12). Some of the observations discussed there will first be briefly repeated here. Both behind an opaque disk, and behind an opening, the wave field has a complicated structure. There are for example always waves along the axis of the shadow cone behind an opaque circular disk (Vol. 1, Fig. 12.13). Behind a circular opening, zones containing waves and wave-free zones alternate along the axis of the blocked-out cone. This was shown by a model experiment, which is reproduced here again in Fig. 21.1. It was explained in Vol. 1, Sect. 12.14 in terms of FRESNEL's zone construction.

We briefly repeat: In Fig. 21.1, the arrows show the observation points P_1 to P_3 which we imagine to lie on the symmetry axis of the wave field. At the point P_2, the opening leaves an *even* number of zones free, namely the two innermost ($m = 1$ and $m = 2$). The elementary waves which they emit cancel each other to a great extent at the point P_2. At point P_3, however, the

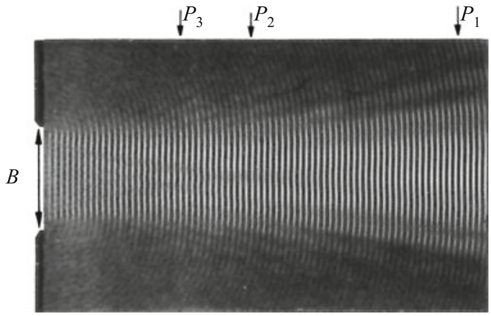


Figure 21.1 A model experiment showing the shadow cast by an opening (Fig. 12.29 in Vol. 1). Suppose that plane waves with a broad wavefront are incident from the left on the opening B. The "cut-out" wave beam extends beyond its parallel geometric shadow boundaries as a result of diffraction. Near the opening, the wave field exhibits a complex structure. *Looking along the direction of the beam, we can see this structure best*. The structure along the beam axis is explained in the text using FRESNEL zones.

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_21

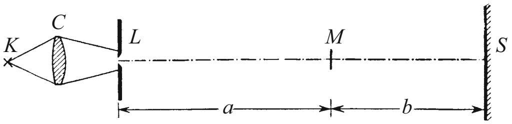


Figure 21.2 A comparison of the shadow of a circular disk M with that of a circular opening of the same size at the same position

opening leaves an *odd* number of zones free, namely the three innermost zones ($m = 1$ to $m = 3$). The elementary waves emitted to point P_3 are conserved. Beyond point P_1, there is no longer any such structure.

With light waves, the situation is similar. Assume that L in Fig. 21.2 is an opening which radiates light waves. It replaces a pointlike wave center as described at the end of Sect. 20.4. The obstacle M which casts a shadow, or a circular opening which limits the beam, is located between L and the observation point. Then for the radius r_m of the m-th zone, we find from Eq. (12.21) in Vol. 1:

$$r_\mathrm{m}^2 = m\lambda \frac{ab}{a+b}\,, \tag{21.1}$$

that is, for $a = b$, $\quad r_\mathrm{m}^2 = m\lambda b/2$; $\quad$ and for $a = \infty$, $\quad r_\mathrm{m}^2 = m\lambda b$.

If the zone radii r_m for light waves in Fig. 21.2 are of the same order of magnitude as those for the waves in the model experiment, then the product λb must take the same values for the light waves as for the waves in the model experiment (Fig. 21.1). However, the wavelength of visible light is more than 1000 times smaller than the wavelength of the waves in the model experiment. As a result, the observation points for the innermost zones ($m = 1, 2, 3 \ldots$) are not, as in Fig. 21.1, only a few centimeters away from the obstacle; instead, they are at distances of many meters. Therefore, in Fig. 21.2, the lengths a and b have values of nearly 20 m. The shadow or diffraction photos made with this setup (Fig. 21.3, a to f) exhibit rather complex diffraction patterns instead of sharp boundaries. They change continually as the distances a and b are varied. In every case, however, they show noticeable differences for circular disks and for circular openings of the same size. Behind the *openings*, we always see only a few rings. At the center of the picture, we can see alternating maxima and minima when the distances a and b are varied. Behind opaque *disks*, the number of rings increases when a and b are decreased, but the center of the pattern always remains brightly illuminated (CHRISTIAN HUYGENS). In the shadow of the disk, the bright spot at the center *persists*; it is called POISSON*'s spot*. It is a point in the shadow of a circular disk, a straight line in the shadow of a rectangular obstacle, etc. POISSON's spot could readily be observed using water waves (Vol. 1, Figs. 12.13 and 12.15. Its origin is discussed there in Sect. 12.14, point 4).

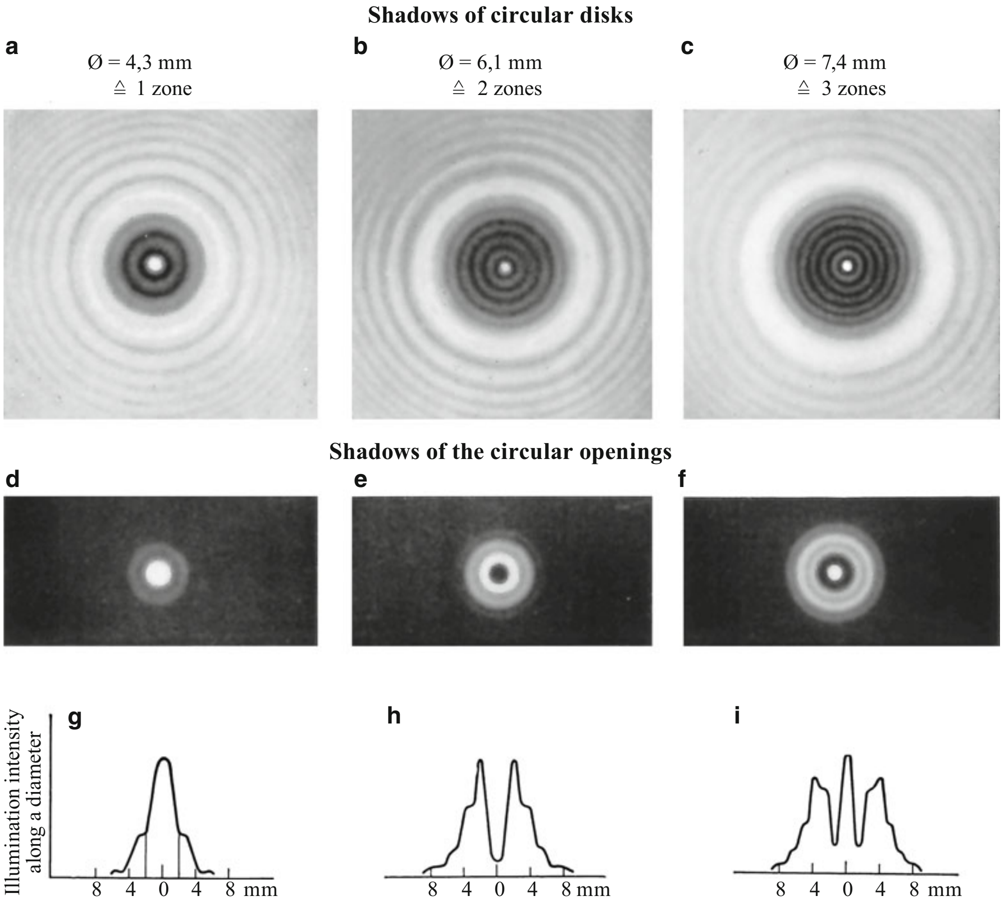


Figure 21.3 The shadows of circular disks and circular openings of the same diameter exhibit very different diffraction patterns. For demonstration experiments, we employ red-filter light. For the photographs (positives), green light with a wavelength of 546 nm was used. The distances a and b were each 17.5 m. The images in the lower series show the distribution of the radiant intensity along a diameter of the patterns in the middle series. Images e and h correspond in Fig. 21.1 to a cross-section perpendicular to the beam axis at the observation point P_2 (**Video 16.3**).

At a distance of $a = b = 11$ km, for red-filter light ($\lambda \approx 650$ nm), the diameter of the central zone is $2r_1 = 12$ cm. That is the size of a small saucer. Such a saucer would thus block out only the central zone from the free wave field. As a result, the shadow pattern of the saucer would look like that in Fig. 21.3a, but its brightest ring would have a diameter of around 50 cm.

For $a = \infty$ and $b = 1$ m, with red-filter light, the diameter of the first zone is $2r_1 = 0.6$ mm. It is thus rather simple (e.g. as in Fig. 16.29) to block out fractions of the first zone with apertures or slits.

Video 16.3:
"Diffraction and coherence" http://tiny.cc/xdggoy.
The **diffraction patterns** shown in the images a–c from circular disks are demonstrated using **thin wires** of different diameters (1.7, 1.0, and 0.2 mm) (from 14:30 min.). FRESNEL's observation mode is employed, and both red- and blue-filter light is used. An explanation of the experimental arrangement is given at the beginning of the video.

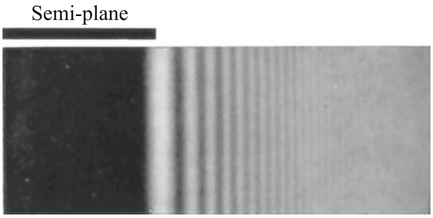


Figure 21.4 The diffraction stripes at the boundary of the shadow of a semi-plane ($a = b = 18$ m, photographic positive, red-filter light) (**Video 16.3**)

Video 16.3: "Diffraction and coherence" http://tiny.cc/xdggoy. Beginning at 12:20 min., the **diffraction stripes behind a semi-plane** are shown in both red- and blue-filter light. An explanation of the experimental arrangement is given at the beginning of the video.

As the diameter increases, both circular disks and circular openings lead to the same limiting case, that of a semi-plane which blocks all the light on one side. The diffraction pattern is shown as a photograph in Fig. 21.4. Every linear shadow boundary looks like this, if the diameter of the light source is sufficiently small.

21.2 Babinet's Theorem

Babinet's theorem is useful for the treatment of diffraction phenomena. A relevant thought experiment is illustrated in Fig. 21.5: From the left, a weakly divergent light beam is incident upon an aperture AB which is several centimeters wide. A light beam emerges to the right. Its boundaries are somewhat fuzzy as a result of diffraction, as is indicated by the shading at the edges of the beam.

Now we draw in a small line segment x. It can represent *either* a small, opaque obstacle *or* a small opening of exactly the same size and shape as the obstacle, in an opaque screen (not shown) that covers the aperture AB.

When x is sufficiently small, the angular deflections of the diffracted waves are large, and the light can penetrate into the regions DD' which were *previously dark*, and illuminate the observation screen there. The diffraction pattern should have the same form for x as an *obstacle* and for x as an *opening*. The reason for this is the following: When the free aperture AB is used *without* x, both diffraction patterns occur simultaneously. Therefore, the wave amplitudes of

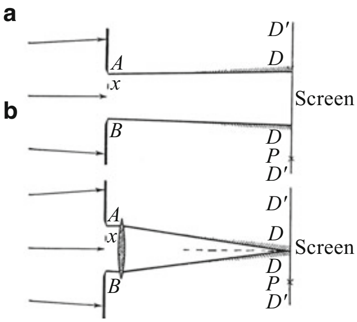


Figure 21.5 Babinet's theorem. Above: Fresnel's observation mode. Below: Fraunhofer's observation mode. If x represents an opaque obstacle, only the free surface area outside it radiates light waves

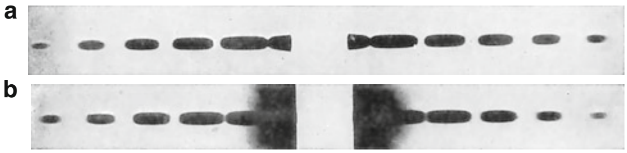


Figure 21.6 a: The diffraction pattern from a wire (obstacle) of 0.5 mm diameter; and b: The diffraction pattern from an equal-sized slit (FRAUNHOFER's observation mode, as in Fig. 21.5b; photographic negative. The center of the image is overexposed in spite of the blocking. The distance to the screen was about 5 m).

the two diffraction patterns must cancel each other exactly at every point in the dark regions DD' at every time. The amplitudes must be equally strong for the obstacle and the opening, and they must have opposite phases ($\delta = 180°$).

This thought experiment leads us to BABINET*'s theorem*. It states that if we insert one after the other an obstacle and an opening of the same size and shape into a wide light beam, and *limit our observations to the regions which were completely dark with only the free light beam* (that is, outside the fuzzy, diffraction-broadened edges of the beam), then in these regions, we will find the *same diffraction pattern* for the obstacle and for the opening.

BABINET's theorem holds both with FRESNEL's and with FRAUNHOFER's observation modes (Vol. 1, Sect. 12.12). With FRESNEL's mode, the diameter of x must usually be made smaller than 0.01 mm. Only then does the diffracted light have a sufficiently large angular deflection, only then can it penetrate into the previously dark regions[1] DD'. One single such small obstacle or opening however gives rise only to an extremely faint diffraction pattern. It requires several thousand of these obstacles or openings x to produce a clearly-visible pattern.

With FRAUNHOFER's observation mode, we consider Fig. 21.5b instead of Fig. 21.5a; then, the free light rays from the aperture AB at an "image point" are concentrated into a narrow band. The dark regions DD' occur on both sides of this band, rather close to the dashed optical axis. As a result, even the *few* diffraction stripes from *large* obstacles or openings x will fall in the dark regions DD'. Then even just *one* opening will produce a readily-visible diffraction pattern.

Figure 21.6a shows, as an example for the validity of BABINET's theorem, the FRAUNHOFER diffraction pattern from a wire. It is the same as that from a slit of the same width, as seen in Fig. 21.6b. For these observations, the center of the diffraction pattern is blocked out by a small screen.

[1] In Fig. 21.3a–c, all the diffraction processes took place *within* the initially-present free light beam. As a result, the decisive precondition for BABINET's theorem was not met, and thus the diffraction patterns from the disk and the opening were quite different.

21.3 Diffraction by Many Equal-Sized and Randomly-Arranged Openings or Particles

With FRAUNHOFER's observation mode (e.g. as in Fig. 21.5b), we make use of a small light source at a large distance on the optical axis. The diffracting opening is placed directly in front of the lens, fulfilling the coherence condition (Eq. (20.1)) for the angle 2ω. The diffraction pattern appears in the focal plane of the lens. We already know from Fig. 17.8 how it will look for a small, circular opening (e.g. of 1.5 mm diameter). *The position of the diffraction pattern is independent of sideways shifts in the position of the opening*. The different segments of the lens always produce a diffraction pattern symmetric to the optical axis. This leads to a conclusion which is of practical importance:

We replace the *single* circular opening by a large number (around 2000) of similar openings of the same size (0.3 mm diameter) in a random arrangement. Then, as seen in Fig. 21.7, we obtain practically the same diffraction pattern as before with only *one* small opening, but it is now visible from a considerable distance and for a large audience. The diffraction patterns of all the openings add with nearly no mutual disturbances. The reason: The light beams from two or more openings can indeed mutually interfere and form additional interference fringes, if they lie within a coherence angle, but the path differences are different for all such combinations. Therefore, the maxima and minima of the additional fringes superpose and cancel, so that on the average, everything remains unchanged, apart from

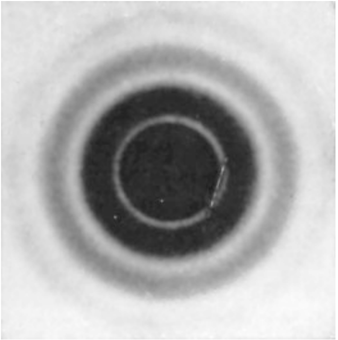


Figure 21.7 The diffraction pattern from a large number of equal-sized, randomly-distributed circular openings (about 2000 distributed over a circular area of 5 cm diameter; the diameter of the individual openings is 0.3 mm. FRAUNHOFER's observation mode, photographic negative image). A small image of the pointlike light source at the center was lost in reproduction. It occurs when the incident radiation is not coherent within the entire angle 2ω subtended by the lens surface **(Video 17.1).**

Video 17.1:
"Resolving power"
http://tiny.cc/9dggoy.
In this video, a mask containing a large number of equal-sized and randomly-arranged holes is placed in front of the lens. See Sect. 17.3.

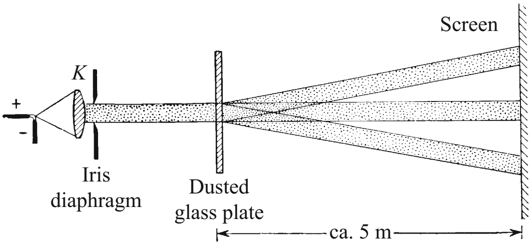


Figure 21.8 Demonstration of the diffraction pattern from many randomly distributed spheres of uniform size using FRESNEL's observation mode. With the dimensions used here, this gives the same result as FRAUNHOFER's observation mode using a lens and convergent beams: The light-wave beams which are deflected to the sides by diffraction are clearly separated from the original beam (the zeroth order diffraction maximum) even without a lens (Sect. 16.8).

a weak *granulation* structure (which is radial in non-monochromatic light). This results from the random arrangement of the openings.

Granulation always occurs in diffraction patterns when coherently-illuminated diffracting objects are arranged randomly. Generally, one observes such granulation subjectively in the diffraction patterns from semi-transparent structures (Sects. 21.4 ff.), for example looking through a frosted glass plate or a "cloudy" windowpane at a small, distant light source. (In both cases, the disordered particles do not have a uniform shape and size. Therefore, the rings are missing.)

In the range of validity of BABINET's theorem, small disks give the same diffraction patterns as openings of the same size. Therefore, we can replace the randomly-distributed openings by randomly-distributed circular disks, and the latter by small spheres of the same diameter: We dust a glass plate with lycopodium seeds, tiny spheres of around 30 μm diameter. With light of wavelength 650 nm (red-filter light), the first diffraction maximum has an angular spacing of about 1.3° from the normal to the plate (the optical axis; cf. Eq. (16.23)). We could thus use FRESNEL's observation mode and project the diffraction rings onto a wall screen. Figure 21.8 shows a suitable experimental arrangement.

21.4 The Rainbow

The little spheres of lycopodium seeds mentioned in the previous section were randomly distributed on a glass plate. Instead of this two-dimensional object, we could use a *three-dimensional*, random distribution of diffraction centers (spheres). Nature provides such a distribution in the form of the fine water droplets in fog and clouds or in a rainstorm. Artificial fog can be readily prepared: We put a little water into a glass bulb and reduce the pressure in the bulb rapidly

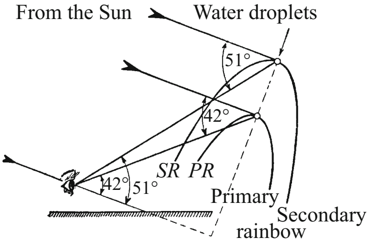


Figure 21.9 Schematic of the primary and the secondary rainbow

with an air pump. This leads to cooling of the air in the bulb, supersaturation of the water vapor and formation of floating droplets. We put a glass bulb filled with artificial fog in place of the dusted glass plate in Fig. 21.8; the diameter of the rings varies with the diameter of the droplets, and the latter increases in the course of time. This can be easily tracked by the decrease in diameter of the diffraction rings.[C21.1]

In a quantitative treatment of this phenomenon, we of course cannot consider the droplets as opaque disks; we must also take into account the waves which pass *through* the spheres. We thus arrive at our first example of diffraction phenomena from *transparent* structures. We start with the facts that are relevant to rainbows (Fig. 21.9):

1. The *primary rainbow* (*PR*) occurs only when the sun is not too high in the sky, at most 42° above the horizon.

2. The central axis of the rainbow lies on a straight line which passes from the sun through the eye of the observer (bottom arrow in Fig. 21.9).

3. Around this axis of symmetry, an arc with an opening angle of ca. 42° is seen; as a rule, it is red on its outer edge, then tinted yellow, green and blue on going inwards. Inside the main arc are several distinct rings which gradually become fainter ("supernumerary" arcs). The sequence of colors in the rainbow reminds us of a spectrum.

4. A second system of rings (arcs), the *secondary rainbow* (*SR*), has an opening angle of 51° around the axis of symmetry. It exhibits the same colors as the primary rainbow (although it is usually fainter), but in the reverse order: red is at its inside edge, then yellow, green etc. going outwards.

The elucidation of these phenomena is found in the *combined effects of diffraction, interference, refraction, dispersion and reflection* from and within the randomly-distributed spherical water droplets. The essential features can most readily be seen with the aid of a model experiment (Fig. 21.10). Here, the water droplets are represented by a thin jet of water, about 1 mm in diameter, which flows from a funnel. Instead of the sun, a line-shaped light source is used (an illuminated slit with a red filter). The screen *S* takes the place of the eye of the observer. On this screen, we see two typical diffraction patterns, i.e. two bands, *PR* and *SR*. With incandescent light, the

C21.1. For more details on the role of diffraction in rainbows, see for example the article by H.M. Nussenzveig, *Scientific American* (1977) (available online at http://www.phys.uwosh.edu/rioux/genphysII/pdf/rainbows.pdf). See also http://www.ams.org/samplings/feature-column/fcarc-rainbows and https://en.m.wikipedia.org/wiki/Rainbow

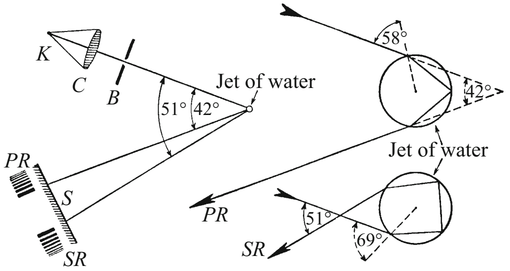
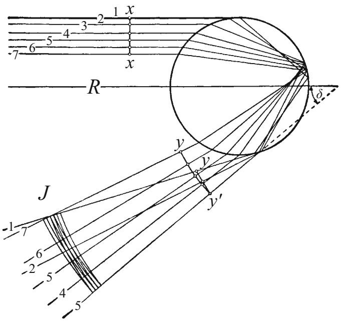


Figure 21.10 A model experiment demonstrating the formation of the rainbow (red-filter light). The jet of water flows downwards through the plane of the page. The screen *S* is set up perpendicular to the plane of the page. It shows the two diffraction patterns, *PR* and *SR*. For subjective observation, we would need a whole "cloud" of parallel jets of water; only then can the interference fringes of various orders from both "rainbows" pass simultaneously into the pupils of the eyes of the observer.

Figure 21.11 Changes in the wavefront by reflection and refraction in a water droplet (calculated for monochromatic light; *xx* before, *yy*′ after passing through the droplet). The beam marked with '*R*' is reflected back onto itself.



well-known superposition of a series of colors is seen. By varying the diameter of the water jet, we can produce many different color series. We can simulate all of the optical phenomena observed in the atmosphere, including the nearly colorless rainbow from extremely fine fog droplets.

We can complement this model experiment, starting with the primary rainbow *PR*, by an elementary calculation. In Fig. 21.11, we consider a parallel light beam which is incident on a water droplet. We first draw several parallel rays 1–7 within this beam, and then, perpendicular to them, a planar wavefront *xx*. We then calculate the paths of the individual rays through the water droplet, applying the law of refraction twice and the law of reflection once. Now comes the essential point: The emerging rays are *concentrated near a certain angle* (the angle of minimum deflection, see Sect. 16.6) at the edge of the diffraction pattern. This results in an overall angle of deflection δ between the incident and the emerging rays of 42° for red light.

We calculate the optical path length for one of the rays between the points x and y (Sect. 16.3). That is, we decompose the segment xy of the beam into individual path elements S_W which pass through water and elements S_A which pass through the air, multiplying the former by the refractive index of water, $n = 1.33$ (for red light; see Table 16.1), and then take the sum to be $nS_\mathrm{W} + S_\mathrm{A} = L$. Then we choose points y along the other rays so that between the points x and y, the optical path lengths of the other rays are also equal to L. Connecting the points y determined in this way gives the new wavefront after passing through the water droplet. Instead of *one planar* wavefront, we now have *two* wavefronts, intersecting at y' and curved. Some of the downstream wavefronts are also drawn in at J, to the left of the calculated wavefronts (yy'). Their lines of intersection yield the diffraction patterns with interference fringes observed at PR in Fig. 21.10. The patterns observed in the secondary rainbow (at SR) are obtained in a corresponding manner with waves (rays) which are reflected *twice* in the interior of the droplets. The point y' lies on the ray with an angle of deflection δ. This angle is 42° for a single reflection and 51° for two reflections.$^{\mathrm{C21.2}}$

C21.2. Larger numbers of reflections can also occur, but the resulting "rainbows" are increasingly faint or located towards the direction of the sun, making them mostly unobservable in the sky. The concentration of scattering paths near a certain angle ("rainbow scattering") is also of great importance in modern nuclear, atomic and molecular physics; see the first reference in Comment C21.1.

21.5 Diffraction by a Step

The first diffraction pattern which we observed was that of a simple slit of width B (Sect. 16.9). Now, we cover half of the slit parallel to its long axis by a *transparent* glass plate, for example a microscope cover glass (of thickness d and refractive index n). Then the covered and the free halves together form a *step*. Its diffraction pattern with monochromatic light is in general asymmetric. It changes periodically when the wavelength is varied continuously, due to dispersion in the glass of the step. We find two symmetric limiting cases: Fig. 21.12, upper left: The path difference $\varDelta = d(n-1)$ is an even multiple of $\lambda/2$; this is the *order-one position*, giving the same pattern as without a step; and Fig. 21.12, lower left: $\varDelta$ is an odd multiple of $\lambda/2$; this is the *order-two position*. We can change $\varDelta$ most conveniently by a slight tipping of the step. *When the maxima approach the central axis more closely, they become stronger; when they move away from it, they become weaker*. These diffraction patterns are explained on the right-hand side of Fig. 21.12 (see also Vol. 1, Sect. 12.13).

21.6 Diffracting Objects with an Amplitude Structure

Both for the case of mechanical waves (Vol. 1) as well as in optics as treated here, we began the discussion of diffraction phenomena with a limiting case: The diffracting objects consisted partly of completely transparent sections and partly of completely opaque sections. A particularly clear example can be found in Vol. 1 (Sects. 12.15 and 12.20,

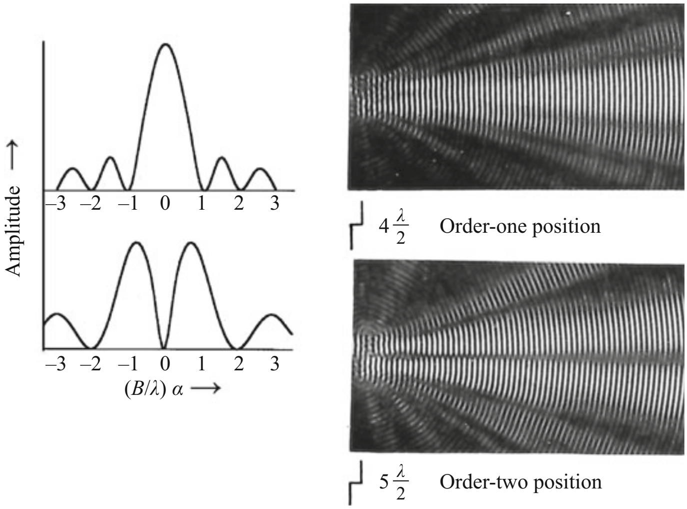


Figure 21.12 *Left*: The two symmetrical diffraction patterns from a step; that is, a slit, half of which is covered by a transparent sheet. Comparison with Fig. 12.33 in Vol. 1 shows that the diffraction pattern for $\Delta = N\lambda$ ($N = 0, 1, 2, \ldots$) is unchanged by adding the step to the slit. For $\Delta = (N + \frac{1}{2})\lambda$, in contrast, there is a minimum at the center of the pattern instead of the primary maximum, and maxima appear near where the first minima were seen before. (B is the slit width, α the angular spacing from the center of the slit; see Vol. 1, Sect. 12.13). *Right*: These results of model experiments are a continuation of Fig. 21.1. The wave centers of the glass images were moved along the step as sketched (**Exercises 21.1 and 21.2**).

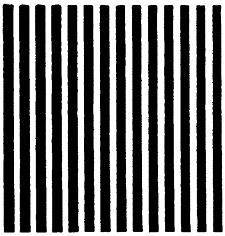


Figure 21.13 A line grating magnified 20-fold. The grating lines are furrows in the surface of a glass plate, filled with an opaque material

point 7): A *line grating*. In such a grating, transparent "slits" alternate with opaque "beams" or "rods". Figure 21.13 shows an example of a line grating. Immediately behind the grating, the amplitudes of the transmitted waves are modulated along a direction transverse to their direction of propagation in a "square-wave" form: The amplitudes alternate sharply between a maximum and a minimum value (Fig. 21.14a). The latter is equal to zero for the special case of completely opaque beams.

More important, however, is a different limiting case: An amplitude modulation as shown in Fig. 21.14b. Immediately behind the grating, the amplitude varies *sinusoidally* around an average value along a direction transverse to the propagation direction of the waves. Gratings

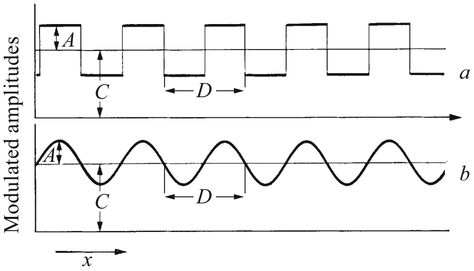


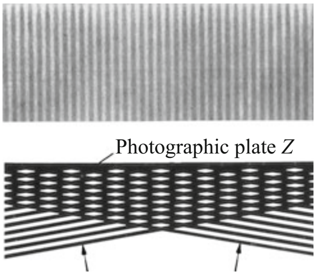
Figure 21.14 Two examples of modulation of the wave amplitude transverse to the propagation direction of the waves, directly behind a line grating. In the upper curve, the constant average value C is equal to A when the curve represents a line grating with completely opaque beams and completely transparent slits. D is the grating constant (length period of the modulation).

which produce this kind of modulation are called *sine-wave gratings*. They can be readily prepared by a photographic method: We generate an interference field with two monochromatic plane-wave light beams which are tilted by a small angle relative to each other (Fig. 21.15). A photographic plate or film is positioned perpendicular to their midline and exposed to the interference pattern. After being developed, it shows the image in the upper part of Fig. 21.14.

Sine-wave gratings have an important property which can readily be shown in demonstration experiments: Symmetrically to the beam which is not deflected (the zeroth diffraction maximum, $m = 0$), they exhibit only two secondary beams of first order ($m = 1$); beams of higher order do not occur.

The method used for fabricating sine-wave gratings – a *combination of interference and photography* – can be experimentally varied in many ways. It can also be used to fabricate line gratings which modulate the amplitudes of transmitted waves as shown in Fig. 21.16 c. This curve can be represented as the superposition of the three amplitude curves sketched below it in the figure. This superposition is not simply formal: The grating acts physically just like three individual sine-wave gratings as shown in the curves d through f. Each of these gratings produces only its two secondary beams of first order

Figure 21.15 Top: A sine-wave line grating (enlarged). Bottom: The procedure for photographically producing the grating shown above, using the interference field of two plane waves which intersect at an angle



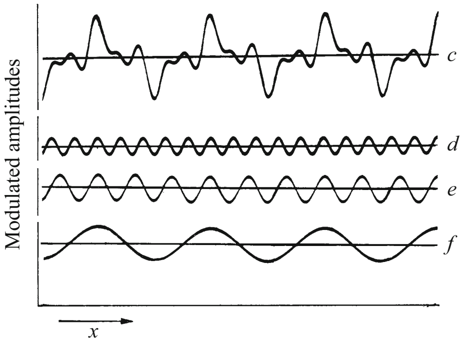


Figure 21.16 The continuation of Fig. 21.14. Curve c shows an additional example of amplitudes modulated along a direction transverse to the direction of propagation of the waves, immediately behind a line grating. Curves d, e, and f are the three FOURIER components of curve c.

($m = 1$) in addition to the incident, practically monochromatic, collimated light beam ($m = 0$) which passes through without deflection (cf. FOURIER analysis). This has a number of applications.

An example:
The grating shown in Fig. 21.13 (a raster grating) modulates the amplitudes of the transmitted waves in a square-wave pattern (Fig. 21.14 a). It acts like a very large number of sine-wave gratings. The grating constants of these "sine-wave gratings" are in the ratios $1{:}\frac{1}{3}{:}\frac{1}{5}\ldots$ (see Vol. 1, Sect. 11.3 and Fig. 11.13). As a result, this grating structure produces a long series of light-wave beams with the orders $m = 1, 3, 5\ldots$ Each of these beams belongs with its order $m = 1$ to one of the individual "sine-wave gratings" (**Exercise 21.3**).

21.7 Gratings with a Phase Structure

We can carry out a continuous transition from grating beams which attenuate the light (Fig. 21.13) to completely *transparent* beams. They need only be distinguished from the "gaps" by their different *refractive index* (G. QUINCKE, 1867). Such transparent structures change only the *phase* of the light which passes through them. In the regions of larger refractive index, the phase changes more than in the regions of smaller refractive index. Thus, we call them for short *phase gratings* or in general *phase structures*.

The diffraction pattern of a phase structure is geometrically no different from that of an amplitude structure of the same form. Differences are seen only in the ratios of the amplitudes and phases between the higher-order and the zeroth-order maxima; the zeroth order may even be completely lacking. An example is given in Fig. 21.17.

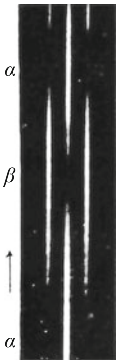


Figure 21.17 The diffraction spectra from a line grating with phase structure, in which the thickness of the "beams" increases in the direction of the arrow. At α, practically only the central zeroth order maximum is present; at β, only the first odd order to the right and the left. In the diffraction pattern from a step (Fig. 21.12), the case at α is the order-one position, and the case at β is the order-two position. To fabricate this grating, a wedge-shaped silver film is deposited onto glass in high vacuum. After the "gaps" have been inscribed, e.g. 5 per millimeter, the silver film is converted by exposure to iodine vapor into transparent AgI.[C21.3]

C21.3. When several emitting areas of width $B/2$ are arranged side by side to form a grating, as for Fig. 21.17, and they emit alternately in phase and with a phase difference of 180° (that is in the order-two position), then the maxima appear unchanged at the same angles and strengths as in Fig. 21.12, below left. Additional maxima are much fainter, and are not shown in Fig. 21.17 (**Exercises 21.4, 21.5**).

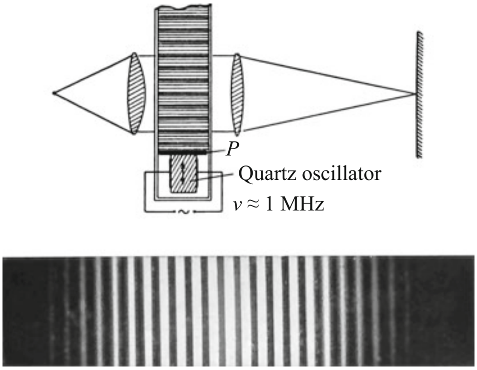


Figure 21.18 DEBYE-SEARS experiment. Top: High-frequency sound waves in a flat liquid-containing basin are used as an optical *phase grating*. The sound waves produce a 3-dimensional layered grating through which the light passes parallel to the layers. It is observed using the FRAUNHOFER observation mode: The propagating sound waves are represented here as an instantaneous image. They are generated by a quartz crystal which oscillates in the direction of the double arrow; it is excited piezoelectrically to oscillations by an electrical oscillator circuit. Below: A diffraction spectrum of this phase grating,[C21.4] photographed in red-filter light.

C21.4. In this layered grating, the density, and with it the refractive index, is sinusoidally varied. The result is thus a sine-wave phase grating, whose grating constant is equal to the wavelength of the sound waves. This produces a diffraction pattern which is similar to that from a line grating with amplitude modulation, if the latter has the same grating constant and narrow gaps in the grating. See L. Bergmann, "*Der Ultraschall*", VDI-Verlag Berlin (1942), 3rd edition, Chap. 2, and especially p. 118. English: see e.g. https://en.wikipedia.org/wiki/Ultrasonic_grating (**Exercise 21.6**).

Differences in the refractive index can result from changes in the density. Sound waves consist of a periodic sequence of regions of increased and of decreased density. Using electrical components, one can readily generate sound waves with wavelengths of order of 1/10 millimeter in liquids, and a narrow basin in which such *sound waves* are propagating can serve as an *optical phase grating* (DEBYE-SEARS, 1932, Fig. 21.18). Observations are carried out using the FRAUNHOFER mode. With it, the diffracting structure can be placed in front of the aperture of the lens used to form an image without changing the position of the diffraction pattern. As a result, it plays no role that the acoustically-produced phase grating is moving past the lens aperture at the velocity of sound. A diffraction spectrum obtained in this manner can be seen in the lower part of Fig. 21.18.

21.8 The Pinhole Camera and the Ring Grating

We can now continue the discussion begun in Sect. 17.2. The central bright spot seen in Fig. 21.3 d as a small white disk and represented graphically in Fig. 21.3 g can serve as the *image point of a pinhole camera*. It uses a circular opening as aperture, which admits only waves from the first FRESNEL zone, the central zone.

> The image point of a pinhole camera is however sharpest when the pinhole allows a region with only 4/5 of the diameter of the central FRESNEL zone to pass through (Fig. 21.19).

The bright spot within the *shadow* of all circular disks (e.g. in Fig. 21.3 a–c) can be used for image formation. Here, we are not limited in our choice of diameter for the disk (POISSON's spot). Figure 21.20 shows an example. There, the disk was replaced by a steel sphere of 4 cm(!) diameter.

More intense images than with only *one* disk or sphere can be obtained by using a large number of concentric, narrow, circularly-symmetric transparent rings which surround a central opaque disk. When the radii of the rings are chosen randomly, only the radiant *power* of the diffracted waves adds at every point along the symmetry axis. The phase relations which would give a large resulting amplitude are lacking, and thus the radiant power, which is proportional to the square of the total amplitude, is lower. Fixed image positions are also lacking. Both can be obtained only if the radii of the rings are chosen to be proportional to the square roots of whole numbers, and only light beams with a small opening angle are used. Then the path lengths which lead through two neighboring rings from an object point to the corresponding image point differ only by integral multiples m of the wavelength λ. The only path differences are $\Delta = m\lambda$,

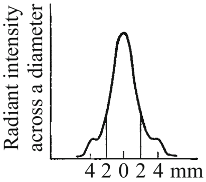


Figure 21.19 The distribution of the radiant intensity in the shadow of a circular hole that allows only 4/5 of the central FRESNEL zone to pass through (the diameter of the hole and the distances are the same as in Fig. 17.9)

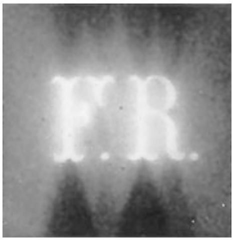


Figure 21.20 An image (actual size) formed using a *steel sphere* as the imaging system (the setup was as shown in Fig. 21.2; the diameter of the sphere is 4 cm, $a = 12$ m, $b = 18$ m). The object is a metal stencil about 7 mm high, in place of the aperture L in the figure.

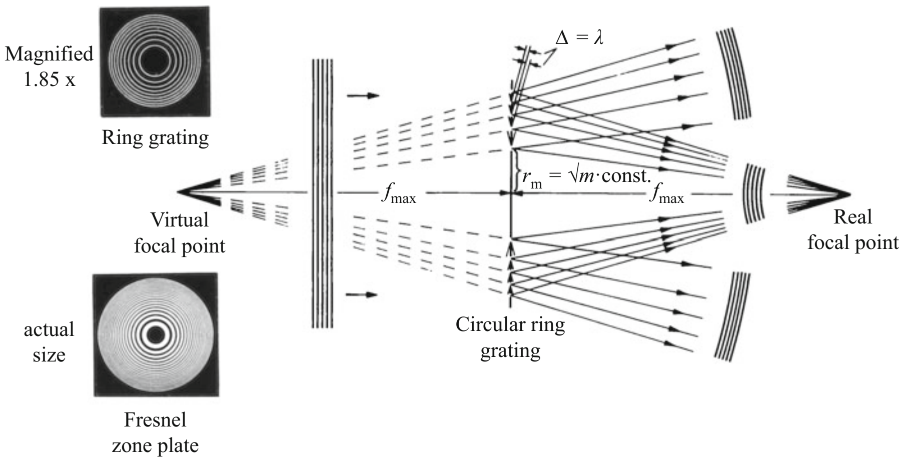


Figure 21.21 Top left: A ring grating which is suitable for use as an imaging system ($f_{\max} \approx 90\,\mathrm{cm}$ for red light). At the right: The simultaneous formation of a real and a virtual focal point. Bottom left: A FRESNEL zone plate with an opaque central disk. It forms ca. 5 real and virtual focal points which can be observed with red light ($f_{\max} \approx 2.7\,\mathrm{m}$).

i.e. all the waves that reach a given image point have the same phase. This is sketched in Fig. 21.21 for the formation of a focal point. Here, the object point is to the left at infinity, and the plane waves coming from it are incident on the ring grating in the direction of the two arrows. The waves which emerge from the rings give rise to both virtual and real focal points. For $\Delta = \lambda$, the longest focal length $f_{\max}$ is found. A simple geometric construction (with $a \to \infty$) yields $f_{\max} = r_1^2/\lambda$. Shorter focal lengths occur at $\Delta = 2\lambda, 3\lambda \ldots$ They obey the relation $f_m = f_{\max}/m$. This type of ring grating produces quite intense images as an imaging system (an example is shown at the right in Fig. 21.22).

In a ring grating, the width of the transparent rings can be increased, at the same time reducing the width of the opaque rings, thus keeping the overall areas of the individual rings roughly the same. In this way, one arrives at a FRESNEL *zone plate* (Vol. 1, Sect. 12.14) with an opaque central disk. It acts in the same manner as its "negative" with a transparent disk at its center. Both distribute the radiant power over many real and virtual images and focal points (this is a result of the sudden transitions between transparent and opaque rings). They are thus inferior to lenses as imaging systems. Only the radial order obtainable with a lens shape allows minimal losses of radiant power in image formation.

In order to demonstrate ring gratings and zone plates as imaging systems with fixed image positions, we can use the arrangement sketched in Fig. 21.22. With it, for example the image shown as a photograph on the right in Fig. 21.22 was obtained. Virtual images can be observed subjectively (with the eye directly behind the grating); but see also the caption of Fig. 21.22.

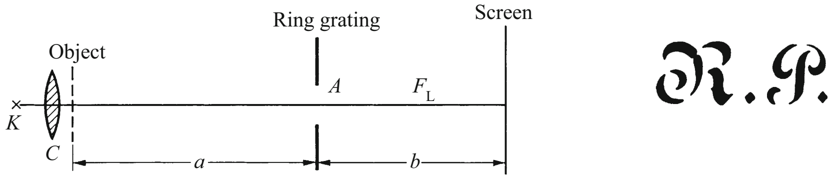


Figure 21.22 The demonstration of ring gratings and zone plates as imaging systems. At the right is the real image of a small stencil, photographed actual size. It was produced by the ring grating shown at the upper left in Fig. 21.21 ($a = b \approx 1.8$ or 0.9 or 0.6 or 0.45 m; K is the crater of an arc lamp, C is the condenser lens for stencils up to ca. 10 cm in diameter). For the subjective observation of virtual images or virtual focal points, it is difficult for the eye not to accommodate to the bright light source. Therefore, it is expedient to replace the lens of the eye by a convex lens, for example at A, with its focal point at F_L. Then we can observe the virtual image on the screen at the right as a real image, while the corresponding real image can be made visible on a screen to the left of F_L. (The justification of these statements can be seen in Fig. 21.21: Imagine that there, directly behind the ring grating, we insert a large convex lens, which encompasses the whole surface of the grating).

21.9 A Ring Grating with Only One Focal Length

Besides the linear line grating with straight-line slits and beams, we have also treated ring gratings in Sect. 21.8. The most well-known of these is a limiting case, the FRESNEL zone plate, with its abrupt transitions between completely transparent and completely opaque rings.

An especially important form of the ring grating can be fabricated photographically using interference. We again use two coherent, monochromatic wave trains of light which intersect, but this time, one has spherical symmetry and the other is a plane wave. A photographic plate is placed in the resulting interference field[2]. Figure 21.23 shows an example.

Ring gratings which are fabricated in this manner by photographing an interference field are called *sine-wave ring gratings*, since two of their characteristics are similar to those of sinusoidally-modulating line gratings (sine-wave gratings; see Sect. 21.6):

1. In a sine-wave line grating, aside from the central collimated light beam which is not deflected (zeroth order), only the two sub-maxima of order $m = 1$ remain. In a sine-wave ring grating, aside from the central maximum, only the two focal or image points belonging to the order $m = 1$ remain (this can be demonstrated with the setup from Fig. 21.22).

[2] If the general clarity of the experimental setup is not important, we can make use of a more modest arrangement: We pass a collimated light beam through a filter practically perpendicular to the planar side of the lens used for the demonstration of NEWTON's rings (see the end of Sect. 20.9), and then we project an image of this surface onto the photographic plate, using a lens with a long focal length.

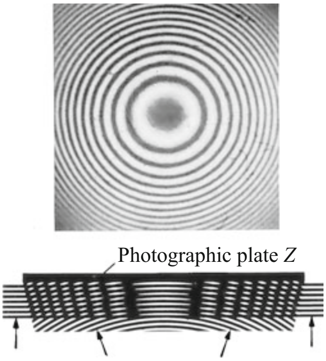


Figure 21.23 Top: A ring grating (magnified 3.4 x) with only one focal length ($f \approx 1.5$ m for red light). Bottom (on a different scale): Its photographic fabrication in an interference field. In order to demonstrate this grating as an imaging system and to show the position of its focal point, we use the experimental setup described in Fig. 21.22.

2. Several sine-wave line gratings with different grating constants can be superposed without losing their individual character. The analogous behavior is found on superposing several sine-wave ring gratings, using long wave trains of monochromatic light.

This important fact can be readily demonstrated today, since monochromatic radiation has become available using lasers$^{C16.4}$ as light sources. The setup sketched at the upper left in Fig. 21.24 suffices. There, a nearly perfectly-collimated monochromatic light beam (i.e. a plane-wave beam) is incident on a photographic plate Z. On the way there, it is scattered by four small particles (of Al_2O_3 or 'alumina'). These particles, carried by a slightly wedge-shaped glass plate, serve as four object points P. The spherical waves which are emitted by the scatterers interfere both with each other and with the primary collimated light beam. The photographic plate records a section through the resulting interference field. It can be seen on

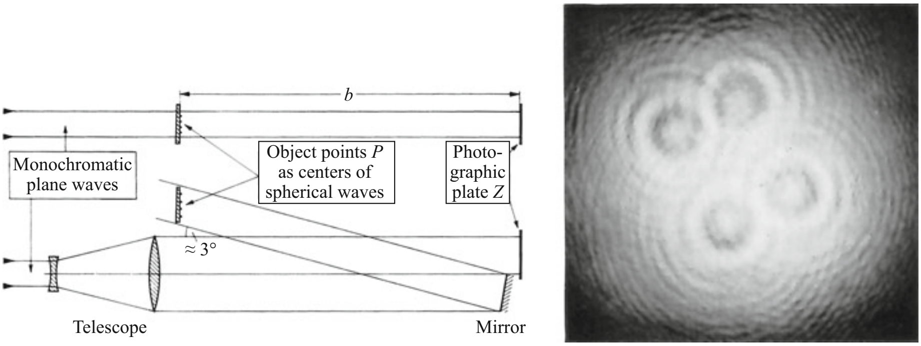

b
Monochromatic plane waves
Object points P as centers of spherical waves
Photographic plate Z
≈ 3°
Telescope
Mirror

Figure 21.24 At the right: Four sine-wave ring gratings which overlap each other (photographic positives, magnified 2x). Upper left: Their photographic production in an interference field ($b \approx 1.3$ m). Lower left: A technically-important variant: The small particles which serve as four object points P or as four sources of spherical secondary waves are illuminated with *incident light*. In order to make room for the mirror, the cross-section of the monochromatic light beam (laser beam) is enlarged using a telescopic optical system (Sect. 18.13). In both sketches, the four object points lie in a single plane. This is not necessary, but it is experimentally convenient.

the right in Fig. 21.24: It shows the superposition of the four sine-wave ring gratings belonging to the four object points. These, as claimed above, should not have lost their individual character if we use long wave trains of monochromatic light for the experiment. The empirical proof: If we remove the four object points and illuminate the developed photographic plate with a monochromatic, collimated light beam, then at a distance b to the *right* of Z on a screen (not shown in Fig. 21.24), we find *real images* of the no-longer-present object points! These are the *real* focal points of the individual sine-wave ring gratings.

The corresponding *virtual* images of the no-longer-present object points can be observed subjectively: We look through Z as if looking through a window. It is usually more convenient to observe the virtual images also as real images projected onto a screen; for this, we proceed as shown in Fig. 21.22.

21.10 Holography

Compact disks and magnetic tape can record sound waves (as a linear sequence) with the correct amplitudes and phases, and reproduce them on demand in the same form as the original, no longer extant sound waves. The experiment described in Sect. 21.9 does the same for light waves (as a two-dimensional recording). That experiment demonstrated the principle of "holography". The image on the plate Z shown at the right in Fig. 21.24 was the *hologram* of an object which consisted of only four object points.

Imagine that these four object points are replaced by the set of all points which make up the monochromatically-illuminated surfaces of some arbitrary objects. Then on plate Z, we will see no recognizable interference patterns from individual sinusoidal ring gratings. To the uninformed eye, the hologram appears to consist of a nearly unstructured grey surface. It is all the more surprising that it can produce a virtual image, for example when we look through the hologram, illuminated with monochromatic light, as if through a window.

A normal photograph shows an impeccably lifelike perspective when viewed with one eye from the correct distance (Sect. 18.16). *A hologram is however superior*: It permits us to
– see an object as if we were gazing at the original itself, and
– see other, previously hidden objects appear when we change our viewing angle! The origin of this superiority is easily understood:

Normal photography makes use of only the first of the two characteristics of a light wave in the recording process on a film or plate, namely its *amplitude*. The square of the amplitude is proportional to the radiant power which arrives at a given "image point". Holography uses in addition the second characteristic of the waves, i.e. their *phases*. It thus increases the amount of "information" stored in the photograph. The frame of reference for these phases in the simplest case is a plane wave which is incident on the photographic film or plate at the same time as the spherical waves that emerge from the object points.

The technical development of holography[3] has undergone rapid progress.[C21.5] We can expect many more applications of holography to come. One point is particularly important and deserves mention: Holography does not require monochromatic light of the same wavelength for the production of holograms and/or for viewing them (analogously to the last paragraph of Sect. 21.9).

C21.5. Holography is today very widespread and finds a variety of applications, e.g. on bank notes and identification cards, in data storage devices and even in archeology. See for example B.K. Buse and E. Soergel, *Physik Journal* **2** (2003), No. 3, p. 37. English: See e.g. https://spie.org/Documents/Publications/00%20STEP%20Module%2010.pdf.

21.11 Visible Imaging of Invisible Objects. The Schlieren Methods

Some objects have neither visible outlines nor visible internal structures. A jet of CO_2 gas streaming from a nozzle into the room air is invisible. A smoothly-polished sheet of glass shows no structure in its interior. These things are not too small or too distant for us to see; the reason for their invisibility is quite different: They modify the visible radiation which passes through them in ways which our eyes cannot register. They do not *reduce the intensity* of the radiation, but rather they change only its *phase*, or at most its *direction* very slightly.

Such invisible objects can be made visible by using a simple trick. We bring them, like an object which can cast a shadow, into the optical axis of a nearly pointlike light source (e.g. the crater of an arc lamp). This "simple schlieren method"[C21.6] is used in Fig. 21.25, left side, to show a jet of CO_2 gas, and on the right side, the internal structure of a sheet of glass. Explanation: Normally, the screen would be illuminated uniformly. However, a light beam which passes through the jet of gas or through the interior of a sheet of glass (its inhomogeneities) is deflected slightly to the sides, in part by *refraction*, in part by *diffraction*. On the screen, therefore, at certain positions some of the light is missing, and these points seem darker. Other positions receive additional radiation, and therefore seem brighter.

C21.6. "Schlieren" refers to the regions which influence incident light due to inhomogeneities in an otherwise homogeneous material.

With this *simple schlieren method*, we have already seen the essentials: *Two groups of light beams are distinguished by their different directions*, one without deflection and the other deflected in part by refraction, in part by diffraction.

[3] The principle was discussed nearly a century ago. Its experimental realization took place, as usual, in a series of steps, first by H. Boersch (1938), then by D. Gabor (from 1948). Among the numerous later works, especially those of E.N. Leith and I. Upanieks (1962) deserve mention. These authors no longer used simply a beam of plane waves as reference frame for the phases, but instead *many* such beams, which are formed by scattering of a monochromatic, collimated beam of light by a frosted-glass disk (cf. 16.8).

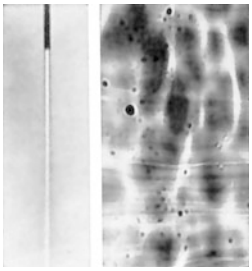


Figure 21.25 Two images, photographed with the simple schlieren method: At left, a jet of CO_2 gas, moving downwards in a laminar flow. At the right: a section of a sheet of glass (a rinsed negative, 9×12 cm). (The distance between the light points to the object and from the object to the screen is several meters in each case; the image is about 1/3 actual size).

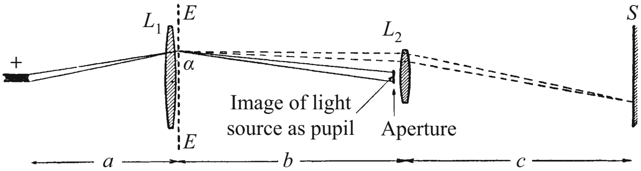


Figure 21.26 Toepler's schlieren method. The object plane EE corresponds to the slide in a slide projector. Only a single partial beam belonging to the object point α in the object plane is sketched with its two bounding rays (imagine that there is a small opening at α). The diameter of the lens L_2 used for imaging must be larger than the entrance pupil. The "aperture" can be either an opaque disk (dark field) or an opening in an opaque screen (bright field). The lens can be tinted in zones outside the pupil, e.g. from inside to outside red, green, etc. Then we see weak schlieren, which deflect the light only slightly, as red, while stronger deflections are seen as green, etc. A numerical example for a *demonstration experiment*: Lens L_1: $f_1 = 1$ m, diameter 12 cm; $a = 1.5$ m, $b = c = 2f_2 = 4$ m.

A second step leads to a more refined method, the Toepler *schlieren method*: This method makes use of either only the deflected beam or only the non-deflected beam. The experimental arrangement (Fig. 21.26) is the same as for the usual projection system (Fig. 18.33), but with one of the following small additional modifications: Either we use a disk to block off the image of the light source (i.e. the entrance pupil at the lens used for image formation, an aperture illuminated from behind), or we block the whole surface of the lens with the exception of the entrance pupil. In the first case, the field of view is in general dark. It is illuminated only where light beams that have been deflected to the side can reach the lens outside the entrance pupil. The structures appear bright against a dark background on the screen: this is called *dark field illumination* (Fig. 21.27). In the second case, no deflected radiation can reach the lens; the structures appear dark against a light background: *bright field illumination*.

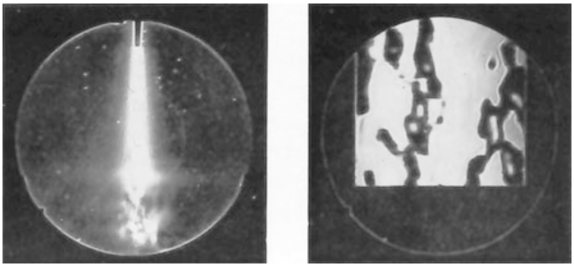


Figure 21.27 Two images photographed with Toepler's schlieren method using dark-field illumination; at left: a turbulent jet of CO_2, flowing downwards. At right: A section of a sheet of glass (around $1/3$ actual size)

21.12 Ernst Abbe's Description of Microscopic Image Formation

When the objects are very small, *deflection of light beams by diffraction* predominates. This holds both for invisible as well as for visible structures. Ernst Abbe treated visible structures in 1873, and his considerations on the role of diffraction in the microscope have proved to be very fruitful.

An experimental setup which we have found to be suitable for demonstrations to a small audience is illustrated in Fig. 21.28. It is analogous to Fig. 21.26. The more important dimensions are indicated, and experimental details can be seen in the caption of Fig. 21.28. The light source should have a *small* cross-section, drawn in the sketch as a square.

In the figure Part B, the object is a large, empty frame β. In the plane Z (position IV, corresponding to the column labelled 'IV' in the tabular lower part of the figure), there is a sharply-focussed image of the light source (shown as a photographic negative), produced by the whole open area of lens L_1 (Sect. 17.2). The light rays which propagate from plane Z to plane S come exclusively from this image of the light source. At plane S, they produce an empty, uniformly illuminated field of view, i.e. the image β' of the empty frame β.

In Part C, the object has an *amplitude* structure: It contains a small, opaque circular disk γ in otherwise open surroundings. In plane Z, in addition to the sharp image of the light source, the *diffraction pattern* from the small disk appears (both shown as photographic negatives). This time, not only light rays from the image of the light source in plane Z arrive at the image plane S, but also the radiation from this diffraction pattern. In the image plane S, both groups of light rays act together and produce the sharp image γ' of the disk, black on a light background (shown as a photographic positive).

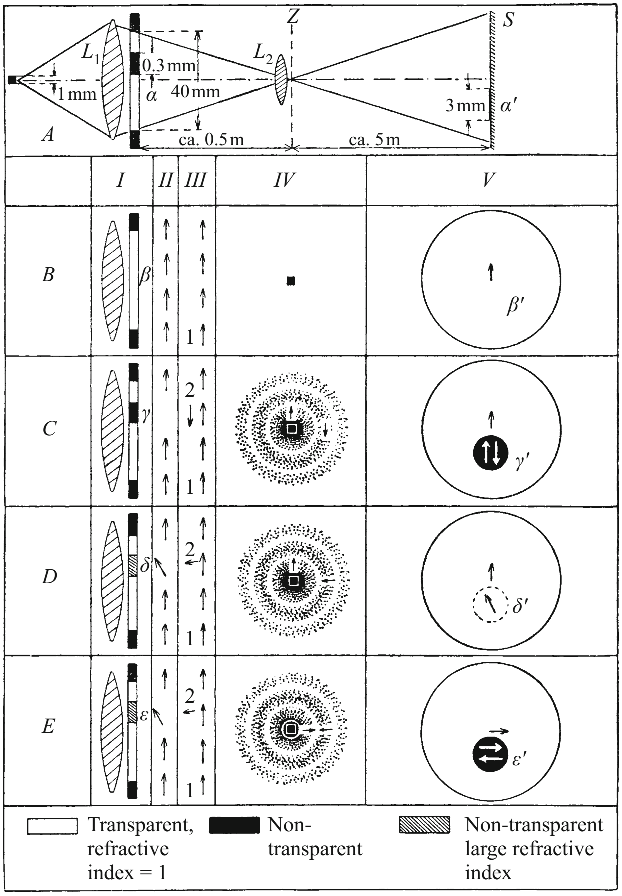


Figure 21.28 Imaging of objects which themselves do not emit light, with amplitude structure and with phase structure. The structure consists of many randomly-arranged circular disks (about 2000 in each case). Their production is described in Fig. 21.29. In the figure, only *one* of these disks in the object (γ) and one in the image (γ') are drawn. The diameter of the diffraction patterns in position *IV* (Abbe's intermediate-image plane) is in reality smaller than the diameter of the imaging lens L_2.

The requirement that *both* of the groups of light rays from plane *Z* are necessary to produce the image in the plane *S* can now be demonstrated by convincing experiments:

1. We place an iris diaphragm in the intermediate-image plane *Z*, then gradually reduce its diameter, and thus, beginning from the outer rim, we block off the diffraction pattern. The result is that the image of the disk γ gradually becomes fainter and fuzzier.

2. In the limiting case, the iris diaphragm allows only the direct light from the light source to pass through. The result is that the image γ'

can no longer be seen; the field of view on the screen S in the image plane is just uniformly illuminated, as in the figure Part B.

3. We now remove the iris diaphragm and block the sharp image of the light source in plane Z with a small disk. The result is that the field of view on the screen S becomes dark. The image γ' of the disk γ appears somewhat fuzzy and bright on a dark background; i.e. we see the amplitude structure of the object imaged with *dark-field illumination* (see the conclusion of Sect. 21.11).

With the aid of these and similar experiments, and referring to the figure, Part C, we describe image formation of an amplitude structure in the following way: After they have passed through the object, we mark the phases of the remaining light rays by vector arrows at position II. The parallel directions of the vectors represent the fact that the rays from individual points reach the image plane S with the same phases. At position III, we decompose the rays formally into two groups:

1. One group of rays from the whole surface of lens L_1, represented by the arrows pointing upwards at 1 (position III). These rays alone produce the sharp image of the light source in plane Z and the uniform illumination of the field of view in plane S. (In Part C, there are also arrows pointing upwards at positions IV and V; these give an arbitrary reference point for the phases of those rays which pass from the small square image of the light source to an image point in plane S).

2. Another group of rays emerging from the object γ, represented by an arrow pointing *downwards* at 2 (position III). These rays produce the diffraction pattern in plane Z and interfere with the rays from the whole surface of the lens in the image plane S, at the image points γ'.

Between these two groups of rays, there is a phase difference of 180° according to BABINET's theorem (Sect. 21.2), represented here by the *oppositely-directed* arrows at positions IV and V. As a result, the rays cancel each other, leaving a dark disk on a light background in position V.

Every change in the groups of rays 1 or 2 will change the interference in the image plane S which is essential for image formation. A correct reproduction of the amplitude structure in the image plane S is thus possible only when both the group 1 of rays from the light source and the group 2 of rays from the diffraction pattern in plane Z can arrive without hindrance at the image plane S.

21.13 Making Invisible Structures Visible Under the Microscope

Most thin sections of organic materials for microscopic examinations in biology and in medicine are transparent and colorless; their chemically distinct structural elements differ in terms of visible light only in their somewhat different *refractive indices*. Often, important structures are just as invisible as the inhomogeneities in a sheet of glass. Most such sections, stated briefly, have practically only *phase structures*. In order to make these structures visible, one has to convert them into *amplitude* structures, i.e. the small differences in their refractive indices must be converted into large differences in their light absorption. To this end, the sections are colored or "stained" with *dyes*, which are taken up selectively by the different structures.

This staining is a chemical intervention and causes considerable deviations from the original state of the living tissues. For this reason, several procedures have been developed for microscopy which can make the phase structures visible without the use of dyes. These procedures are best explained using the terminology of ABBE (Sect. 21.12). We continue the series of images from Fig. 21.28 and show as Part *D* an object δ with *phase* structure: The opaque disk γ in the figure Part *C* has been replaced by a transparent disk δ. It differs from its surroundings only by having a somewhat larger refractive index. The light which passes through this disk arrives at the image plane *S* with a phase delay. This is represented in positions *II* and *V* by a *rotation* of the vectors in a counter-clockwise direction. In the intermediate-image plane *Z*, the image of the light source is surrounded by the diffraction pattern of the disk, just as in the figure Part *C*. Light from both propagates to the image plane *S*. Its superposition there leaves the illumination of the image δ' just as strong as in its surroundings, so that the phase structure remains invisible. In order to make it visible, some sort of intervention in one of the two groups of light waves leaving plane *Z* is necessary, e.g. a partial blocking of the diffraction pattern, or blocking of the image of the light source. In either case, the disk will become visible at the location of the image δ'.

A partial blocking of the diffraction pattern in the intermediate-image plane *Z* can be obtained in a particularly simple manner by using an *oblique* illumination. We shift the light source to one side and thus at the same time, the diffraction pattern shifts (in the opposite direction). Then we can use the lens mount of the objective lens L_2 to cut off an outer portion of the diffraction pattern.

Such a procedure is however a rather rough intervention. A finer method, which produces much better results, is that published by F. ZERNIKE[C21.7] in 1932, the *phase-contrast method*. We describe it in terms of its most important application, to *small* differences in the refractive indices. Parts *D* and *E* of Fig. 21.28 serve to illustrate the method. In these figures, as mentioned above, the phase vector behind the disk δ (position *II*) was *rotated* slightly in a counterclockwise sense. In position *III*, the phase vectors are again decomposed *formally* into two components. The components denoted as 1 produce the image of the *light source* in the intermediate-image plane *Z* (position *IV*, vertical arrow). The component denoted as 2 produces the diffraction pattern in plane *Z* (position *IV*, nearly horizontal arrow). At position *IV*, when there is a phase structure, the

[C21.7] FRITS ZERNIKE, 1888–1966, Dutch physicist. For his invention of the phase-contrast microscope, he was awarded the Nobel prize in 1953.

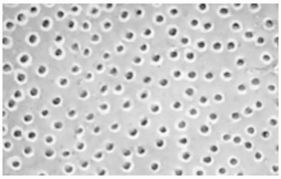


Figure 21.29 A section of a phase structure (magnified about 3x) which is invisible to the eye without making use of some special trick. The small circular disks are made of LiF, embedded in Canada balsam. The LiF was evaporated in high vacuum. As stencil for the evaporation, the same mask was used with which the diffraction pattern in Fig. 21.7 was obtained; it contains around 2000 randomly-arranged holes of 0.3 mm diameter.

two arrows 1 and 2 form an angle of only about 90°, while for the *amplitude* structure, their mutual angle is 180°, i.e. they are directed *oppositely* to each other. We can, however, retroactively increase the phase angle of around 90° to an angle of around 180°. To do this, in the figure Part *E*, the image of the light source is blocked by a small transparent disk (the white circle in position *IV*), which delays the light by 90° (a "quarter-wave plate"). With their path difference now increased to around 180°, the two vectors 1 and 2 act in the image plane as their *difference*, and this produces a good contrast relative to the surroundings (Fig. 21.29). It can be improved still further by making the quarter-wave plate weakly absorbing, and thereby adjusting the lengths of vectors 1 and 2 to be more nearly equal. As the absorption is increased, we follow a continuous transition to a normal microscope with dark-field illumination.

21.14 X-ray Diffraction

The wavelengths of X-rays lie in the range from about 10^{-13} m to $5 \cdot 10^{-8}$ m. The fundamental experiments on diffraction and interference can be carried out just as well with X-rays as with visible light. We mention for example diffraction by a single slit (slit width 5 to $10\,\mu$m),C21.8 and, in particular, the recording of diffraction spectra using typical optical reflection gratings made of metal or glass. These are employed with nearly grazing incidence, since the ruling on the gratings is sufficiently fine only when it is strongly foreshortened by a very flat incidence angle (see Sect. 22.6).

C21.8. As an historical note, we mention here the following relevant publication by the author: Robert Pohl, "*Die Physik der Röntgenstrahlen*", Vieweg and Son, Braunschweig (1912), Chaps. 2 and 9.

For short-wavelength X-rays ($\lambda < 2 \cdot 10^{-9}$ m), mechanically-ruled gratings play only a limited role. Instead, one makes use of the 3-dimensional gratings provided by nature in the form of crystal lattices, as first suggested by M. v. Laue in 1912. Such gratings consist of a three-dimensional series of lattice points or lattice planes, a structure with three (in general different) grating constants (spacings of the lattice planes), D', D'', and D'''. Starting from the condition for

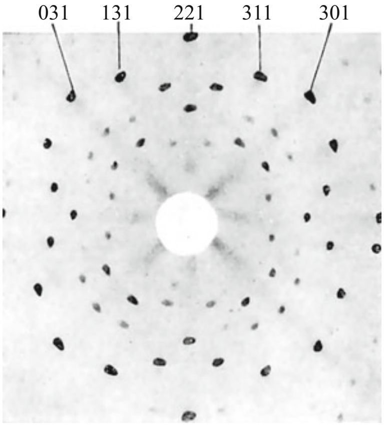


Figure 21.30 A Laue diagram obtained from an NaCl crystal. The number triplets refer to the orders m', m'' and m''' in Eq. (21.2), corresponding to the three spatial directions.

constructive interference by waves reflected from a simple layered structure (see Sect. 22.7, and also Vol. 1, Sect. 12.15 and Fig. 12.41c):

$$\sin\gamma_\mathrm{m} = \frac{m\lambda}{2D} \tag{21.2}$$

(m is the order, and γ is the "grazing angle"),

it can be seen that this condition must be met in a three-dimensional grating for all three spatial directions in order to observe interference maxima.[C21.9] One obtains a pattern of spots, as shown for example in the Laue diagram in Fig. 21.30. Figure 21.31 shows an historical experimental setup.

X-ray diffraction has attained its primary importance today in the field of crystallography. It is the most important method for investigating crystal structures, currently in particular those of large molecular crystals such as proteins. One uses X-rays of known wavelengths and determines not only the *positions* of the interference maxima, but also the *distribution* of the radiant power over spectra of different or-

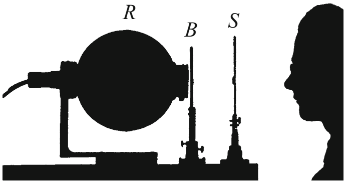


Figure 21.31 A convenient setup for the individual observation of Laue diagrams.[C21.10] R is the X-ray source, with a tungsten anticathode at a voltage of $6 \cdot 10^4$ V; B is a lead shield with the crystal under investigation at its center, in front of an aperture of 2.5 mm diameter; S is a fluorescent screen with a metal disk at its center to block the direct X-ray beam.

C21.9. A further historical note: Up to the 10th edition of "*Optik und Atomphysik*" (1958), a discussion of diffraction from layers and from three-dimensional point gratings was given. For a detailed treatment in English, see e.g. A. Khavasi, K. Mehrany, and B. Rashidian, "Three-dimensional diffraction analysis of gratings based on Legendre expansion of electromagnetic fields", *Journal of the Optical Society of America* **B24** (2007), pp. 2676–2685 (online at https://www.osapublishing.org/josab/abstract.cfm?uri=josab-24-10-2676)

C21.10. Experimenting in this manner with X-rays in the lecture room is certainly no longer permitted today! The observer at the right in the picture is H.U. Harten (Dr. rer. nat. 1949). A detailed description of this experiment by W. Martienssen can be found in the *video* **"Simplicity is the mark of truth"** (at ca. 16 min.).

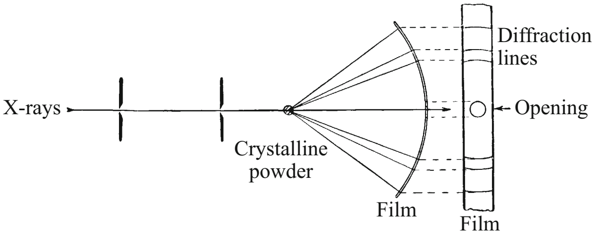


Figure 21.32 The experimental setup of P. Debye and P. Scherrer

ders. From this distribution, one can compute backwards to obtain the detailed structures of the elementary lattice components.

This important method of crystallographic investigation is by no means limited to large single crystal samples. Arbitrarily fine crystalline powder can also be used (P. Debye and P. Scherrer, 1916). As shown in Fig. 21.32, a narrow, collimated beam of X-rays (cross-sectional area around 1 mm^2) is passed through the powder sample and the resulting diffraction pattern is captured by photographic film bent into a circular form (or by a digital-electronic position-sensitive detector). The diffraction pattern consists of a system of concentric

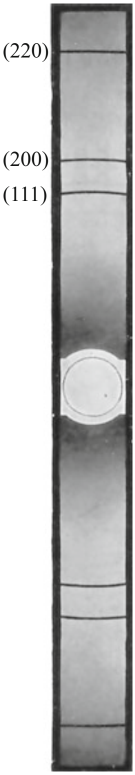


Figure 21.33 A supplement to Fig. 21.32: The K_α radiation from copper ($\lambda = 1.539 \cdot 10^{-10}\,\mathrm{m} = 0.1539\,\mathrm{nm}$) is reflected by three different families of lattice planes in a microcrystalline, well-tempered nickel wire (substituted for Ni powder). The radius of curvature r of the film was 121 mm and its length was πr. (Grating constant $D = 0.3518\,\mathrm{nm}$; the numbers in parentheses are the indices of the reflecting lattice planes. A circular hole for the primary beam is at the center of the film.)

rings (Fig. 21.33). Their interpretation is simple: In a powder sample, the orientation of the crystallites is random. All of the lattice planes on which the X-rays are incident at a grazing angle (Eq. (21.2)) reflect this incident radiation. With large-grained powder, one can clearly see that the rings are composed of a series of individual spots.

Exercises

21.1 How does the diffraction pattern of a slit change
a) if it is completely covered by a transparent glass plate (a microscope cover glass $d = 175\,\mu\text{m}$ thick, with a refractive index of $n = 1.50$);
b) if only one half of the slit is covered with the glass, parallel to its long axis?
c) By what angle α must the glass be tipped to pass from an order-one position to an order-two position? (The wavelength of the light is $\lambda = 600\,\text{nm}$ and the influence of refraction in the glass on the phase difference is negligible.) (Sect. 21.5)

21.2 Derive the diffraction pattern in the order-two position in Fig. 21.12, below left, making use of the graphic construction in Fig. 12.34 of Vol 1.
(Sect. 21.5)

21.3 Why must the slit width B of a high-resolution grating spectrometer be small in comparison to the grating constant D? To answer this question, consider up to which order m one can observe the interference maxima in the intense primary maximum of the diffraction pattern from the entrance slit,
a) for $D = 2B$ (Fig. 21.13), and
b) for $D = 20\,B$, i.e. ($B \ll D$).
(Sect. 21.6, 22.2)

21.4 From the line grating with an amplitude structure as shown in Fig. 21.13 (grating constant $D = 2B$, where B is the slit width), we can construct a phase grating if the opaque lines ruled on the grating are made of transparent material. At which angles do the principal interference maxima lie (cf. Fig. 22.6), if the transparent lines delay the phase of the light waves by
a) 360° (order-one position) and b) 180° (order-two position)?
(Sect. 21.7)

21.5 From the diffraction pattern shown in Fig. 21.17, determine the ratio of the width L of the ruled lines on the grating to the grating constant D. (Sect. 21.7)

21.6 In the DEBYE-SEARS experiment (Fig. 21.18), with a sound wave of frequency $\nu = 10^6$ Hz in xylol, and employing red-filter light ($\lambda = 700\,\text{nm}$), one observes diffraction maxima of m-th order at the angles $\alpha_\text{m} = m \cdot 8 \cdot 10^{-4}$. The grating constant of the sinusoidal phase grating is equal to the wavelength of the sound waves. Determine the velocity of sound c in the liquid. (Sect. 21.7)

Optical Spectrometers 22

22.1 Prism Spectrometers and Their Resolving Power

Optical spectroscopy apparatus are currently technically highly developed and are commercially available, mainly as very convenient recording instruments for a wide range of wavelength ranges[1]. Modern physics needs to consider only the basic questions of the design and operation of spectroscopy apparatus.

Elementary treatments begin with *prism spectrometers*. This name is unfortunate; the prism shape is not essential, but rather the *dispersion* of the material, that is the dependence of the refractive index n for monochromatic (single-frequency) waves on their wavelengths λ. For this reason, we produced the first demonstration of a continuous light spectrum in Sect. 16.10 not with a prism, but instead using a thick glass block with *plane-parallel* surfaces. We also used this same glass block to photograph the line spectrum shown in Fig. 22.1, which consists of the spectral lines from a low-pressure Hg-arc lamp.

A prism spectrometer based on dispersion is sketched in Fig. 22.2. If its entrance slit S_0 is illuminated with monochromatic light, the slit will be *imaged* on the observation screen. The narrower the slit S_0, the sharper will be its image. But there is a limit to this; if it is exceeded, we no longer obtain an *image* of the slit on the screen, but instead the *diffraction pattern* of a light beam which has been restricted

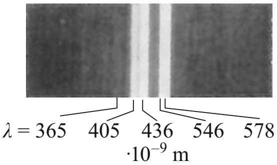


Figure 22.1 A line spectrum, obtained not with a prism, but instead using a glass block with plane-parallel surfaces. It shows the spectrum from a low-pressure mercury-arc lamp, photographed at position b in Fig. 16.30. Compare it with the photographic spectrum reproduced in Fig. 27.3.

[1] Recording spectrometers often register spectra not as a function of the wavelength λ, but rather as a function of its reciprocal $1/\lambda = \nu/c$, proportional to the light frequency or energy. This quantity is often (erroneously) called the *wave number*. The reciprocal of a length is not a number!

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_22

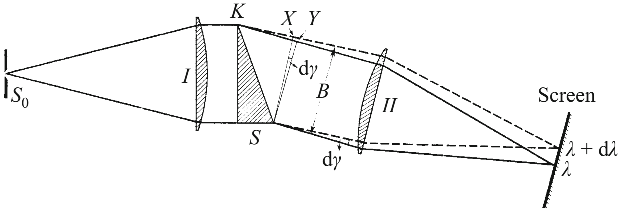


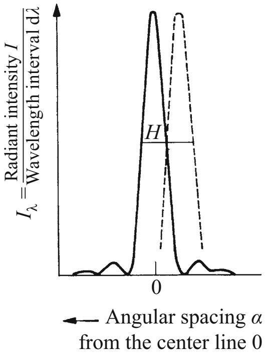
Figure 22.2 Schematic of a prism spectrometer which makes use of dispersion. Light of wavelength λ has a refractive index n. Light of wavelength $(\lambda + \mathrm{d}\lambda)$, which is refracted by a smaller angle, has a smaller refractive index $n + \mathrm{d}n$ ($\mathrm{d}n$ is thus negative). (For individual observations with the eye, a transparent wavelength scale is used instead of the screen, and it is observed from the right using a magnifying lens (ocular).) The whole surface of the prism at K must be illuminated. (K, X, and Y denote points along the path of the upper light beam in the figure). In determining the irradiation intensity on the screen (W/m^2, Eq. (19.4)), the only decisive factor is the numerical aperture of the light beam to the right of lens II (that is, the sine of the opening angle of the beam). Therefore, for the observation of line spectra, it is not necessary that the lens I have a short focal length, nor does the slit S_0 need to be set at an inconveniently narrow width. When combined in a tube, S_0 and I are called a *collimator*.

to the width B. This pattern lies in the focal plane of the lens II, and is thus observed in the FRAUNHOFER mode. The formation of such diffraction patterns was described quantitatively in Sects. 12.12 and 12.13 of Vol. 1, and experimentally demonstrated there using undamped, i.e. monochromatic sound waves with a wavelength of ca. 1 cm. For light, a similar experiment is shown in Fig. 16.29. We imagine that the slit S_0 in Fig. 22.2 is illuminated with two wavelengths, λ and $(\lambda + \mathrm{d}\lambda)$,[C22.1] and is closed down to such a narrow width that instead of two *images* on the wavelength scale, we see two *diffraction patterns*. This is sketched in Fig. 22.3 for a limiting case: The halfwidths H of the two diffraction patterns just touch each other, so that the central maximum of the one pattern falls in the first minimum of the other. With this spacing, the eye is just able to distinguish the two patterns as separate; they are barely *resolved*. (This condition is known as the RAYLEIGH criterion).

In Fig. 22.2, both light beams, the one drawn with solid lines and the one with dashed lines, have practically the same widths B. For light of wavelength λ, the optical path passes through the base of the prism S: $KY = S \cdot n$, where KY is the length of the line segment between the points K and Y. For light of wavelength $(\lambda + \mathrm{d}\lambda)$, it is $KX = S \cdot (n + \mathrm{d}n)$. To a good approximation, $KX - KY = S \cdot \mathrm{d}n$, so that the wavefronts of the two light beams are tilted by different angles. The difference $\mathrm{d}\gamma$ between these two angles is $\mathrm{d}\gamma = S \cdot \mathrm{d}n/B$. On the screen, two diffraction patterns appear, belonging to the two light beams as sketched in Fig. 22.3. We observe $\mathrm{d}\gamma < 0$, that is, $\mathrm{d}n < 0$. A dispersion $n(\lambda)$ (Sect. 16.10) with $\mathrm{d}n/\mathrm{d}\lambda < 0$ is called *normal dispersion*. For the solid-line diffraction pattern, the

[C22.1] The quantities preceded by the letter d in this chapter are not differentials, but rather small differences in the corresponding quantity, in this case the wavelength λ. The notation $\Delta\lambda$ will be introduced later for the "usable wavelength range".

Figure 22.3 The resolving power of a prism spectrometer. The observer looks from the right, that is *into* the direction of the oncoming light, towards the screen sketched in Fig. 22.2. The width of a diffraction pattern line at the points where its ordinate values on both sides have decreased to half of their maximum value is called its *halfwidth* H.



first *minima* on both sides of the center line are marked. As seen from the midpoint of lens *II*, they are separated from the center line by the *small* angle $\alpha = \lambda/B$ (according to Eq. (16.23)). The dashed diffraction pattern, which belongs to $(\lambda + \mathrm{d}\lambda)$, has practically the same shape, and it is clearly separated from the solid-line pattern when $\mathrm{d}\gamma = -\alpha$, or $S \cdot \mathrm{d}n/B = -\lambda/B$ (Rayleigh criterion). This leads us to the *resolving power* of the prism:

$$\frac{\lambda}{\mathrm{d}\lambda} = -S\,\frac{\mathrm{d}n}{\mathrm{d}\lambda}\,. \tag{22.1}$$

In words: The *resolving power of a completely illuminated prism is determined solely by its base length S and by the dispersion* $\mathrm{d}n/\mathrm{d}\lambda$ *of the prism material* (as shown in Fig. 22.2). (In the case of *anomalous dispersion*, that is when $\mathrm{d}n/\mathrm{d}\lambda > 0$, the minus sign should be removed).

Numerical example: Assume that a prism with a base length of $S = 1\,\mathrm{cm}$ is made of flint glass, with a dispersion in the "yellow" spectral range of $|\mathrm{d}n/\mathrm{d}\lambda| = 10^3/\mathrm{cm}$. Its resolving power is then $\lambda/\mathrm{d}\lambda = S \cdot |\mathrm{d}n/\mathrm{d}\lambda| = 1\ \mathrm{cm} \cdot 10^3/\mathrm{cm} = 10^3$. The prism can just separate the two sodium D lines. Their wavelengths are $\lambda_{\mathrm{D}_1} = 589.0\,\mathrm{nm}$ and $\lambda_{\mathrm{D}_2} = 589.6\,\mathrm{nm}$. Their spectral separation thus requires

$$\frac{\lambda}{\mathrm{d}\lambda} = \frac{589\ \mathrm{nm}}{0.6\ \mathrm{nm}} \approx 10^3\,.$$

Using three prisms one after the other, made of the same material, each with a base length of 10 cm, we could attain a resolving power of $\lambda/\mathrm{d}\lambda = 3 \cdot 10^4$.

22.2 Grating Spectrometers and Their Resolving Power

Grating spectrometers can be used in all wavelength ranges, at least from the region of X-rays out to electromagnetic waves of several centimeter wavelength. For visible light, the prism in Fig. 22.2 is replaced by a line grating (Fig. 22.4). Line gratings (Fig. 21.13) were

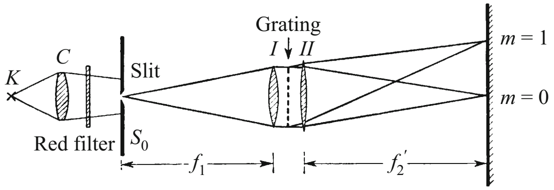


Figure 22.4 A grating spectrometer (J. Fraunhofer, 1821). *m* is the order of the diffraction maximum or minimum. For demonstration experiments, the right-hand lens *II* is usually left off and the screen is placed at a distance of several meters (cf. Fig. 16.26). The boundaries of the light beams are drawn in for the central maximum ($m = 0$) and a spectral line of first order ($m = 1$). For subjective observation, we replace lens *II* by the objective and the screen by the focal plane of a telescope (**Exercise 22.1**).

treated in detail in Vol. 1 (Sect. 12.15 and Fig. 12.65), based on extensions of Young's interference experiment. Gratings sort spectral lines of the *same* wavelength λ according to their *orders m*.

For the angular spacing α of a spectral line of mth order to the symmetry plane perpendicular to the plane of the grating, we have found:

$$\sin \alpha_\mathrm{m} = \frac{m\lambda}{D} \tag{22.2}$$

(D is the spacing of two neighboring openings or wave sources, the *length period* or "grating constant"; $m\lambda$ is the path difference of the wave trains from two neighboring openings).

In making the transition from Young's interference experiment, which makes use of *two* light beams, to a grating with N light-wave beams, the interference maxima *become sharper while retaining their positions*. In this process, between each two neighboring maxima, $(N-2)$ submaxima appear (Vol. 1, Fig. 12.40). This is demonstrated in Fig. 22.5 for light waves. In the bottom image, with $N \approx 250$, the submaxima have practically disappeared and the principal maxima have become very narrow diffraction patterns (when the width of the slit S_0 is sufficiently small).

With N grating lines, a spectral line of mth order is separated from the neighboring order $(m + 1)$ by $(N - 2)$ submaxima, and thus by $(N - 1)$ minima (Fig. 22.6). The spectral line of mth order occurs for a path difference of $m\lambda$ between two neighboring wave trains. For the next following spectral line of order $(m + 1)$, this path difference has increased by a *whole* wavelength λ. It follows that it has increased for the next minimum (γ), which follows the line of mth order, by only a *fraction* of λ, namely from $m\lambda$ to $(m\lambda + \lambda/N)$. Now, we want to be able to distinguish a spectral line of order m and wavelength $(\lambda + \mathrm{d}\lambda)$ from the spectral line of order m and wavelength λ. For that purpose, the line of wavelength $(\lambda + \mathrm{d}\lambda)$ must fall at least in (or outside) the first minimum (γ) adjacent to the spectral line of

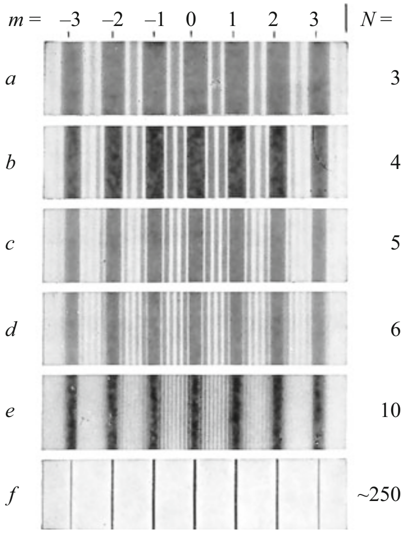


Figure 22.5 The interference pattern of a line grating, showing its dependence on the number of interfering light beams (the number *N* of grating lines). *m* is the order. For the figure Parts *a*-*e*, red filter light suffices; for *f*, the light from a sodium-vapor lamp was used (photographic negative).

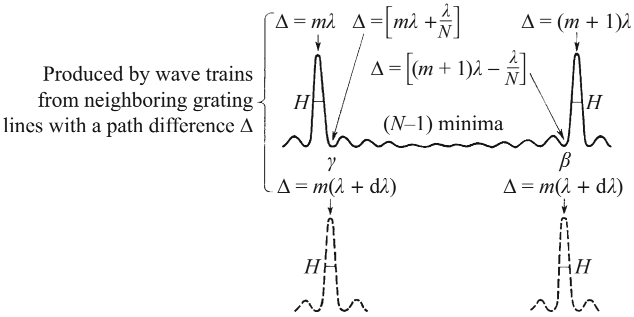


Figure 22.6 The resolving power and the usable wavelength range Δλ (see Sect. 22.5) of a grating spectrometer. In the interest of clarity, the spectral lines (solid and dashed lines) are not drawn beside each other, as in Fig. 22.3, but rather one above the other. The observer is looking at the image plane of the lens *II* in Fig. 22.4 along the direction of the light propagation. If *N*, i.e. the number of grating lines, is increased in this figure, then the submaxima already present move together on both sides of the neighboring principal maxima; the ratio of their heights to that of the principal maximum remains unchanged. At the center between two principal maxima, new submaxima are formed, and they become smaller and smaller in height. Compare Sect. 12.15 in Vol. 1 (**Video 20.1**).

Video 20.1: "Interference"
http://tiny.cc/9eggoy.
The second part of the video shows **interference patterns from different gratings** (Grating constants: 20 and 10 μm) using light from a laser (λ = 633 nm, beam diameter 3 mm) and the lecture-room wall as observation screen.

wavelength λ (RAYLEIGH criterion). With this condition, we obtain

$$m(\lambda + \mathrm{d}\lambda) = m\lambda + \frac{\lambda}{N} \quad \text{or} \quad \frac{\lambda}{\mathrm{d}\lambda} = Nm\,. \qquad (22.3)$$

In words: The *resolving power* $\lambda/\mathrm{d}\lambda = \nu/\mathrm{d}\nu$ of a grating in the first-order spectrum is equal to the *number N of grating lines*. For spectral lines of higher orders m, it increases proportionally to m. Numerical examples for the resolving power of commonly-used gratings will be given later in Sect. 22.6.

22.3 The Line Shapes and Halfwidths of Spectral Lines

An almost uncountable number of articles have been written on the experimental investigation of line spectra. In many cases, the results obtained with spectrometers are presented graphically and the intensities of the radiation are indicated by the width of the lines drawn. For some aspects of the investigations, this method of representation is not adequate. The frequency and the wavelength alone are not sufficient. An additional quantity that must be determined is the *shape* of the spectral lines, characterized by their *halfwidths*.

A spectral line thus has two characteristic parameters; first, its center frequency ν_0, and second, its halfwidth H. The latter is the difference $\Delta\nu$ between the two frequencies at which the ordinate (line strength or amplitude) has decreased to half its maximum value. When the spectrum is plotted against the wavelength λ, sometimes the corresponding *wavelength difference* $\Delta\lambda = H_\lambda$ is defined as the halfwidth of the line.

As an example, Fig. 22.7 shows the line spectrum of a high-pressure mercury-vapor lamp, registered with a prism spectrometer. In this spectrum, for example, the "blue" spectral line (at $\lambda = 436$ nm) has a large halfwidth, $H_\lambda = \lambda/60$.

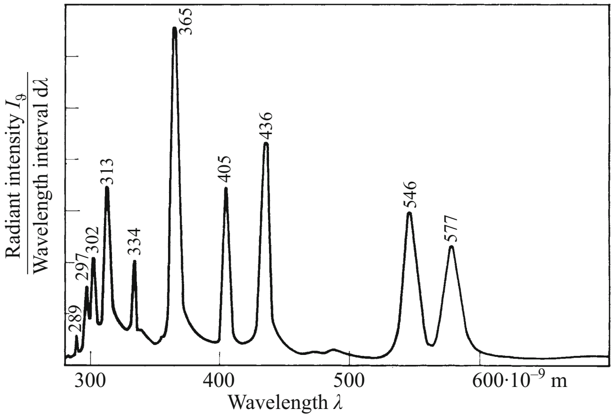


Figure 22.7 The spectral distribution of the radiant intensity from a high-pressure mercury-vapor lamp (line spectrum), with very broad spectral lines, registered using a prism spectrometer. This instrument has a resolving power of $\lambda/d\lambda = 6000$, and thus has no measurable influence on the shapes of the spectral lines (see also Fig. 27.3).

22.4 Spectrometers and Incandescent Light

Incandescent or "natural" light does *not* consist of monochromatic waves from a wide frequency region, but by using *spectrometers*[2], we can produce *quasi-monochromatic* waves from incandescent light. (For this application, the apparatus is often referred to as a "monochromator"). Its function can be seen *qualitatively* directly by considering prism spectrometers which employ dispersion: Every dispersion converts non-periodic processes into periodic processes. Such a reforming of wave groups can be seen on the water surface of a pond into which a large stone has been dropped: At first, we see a non-periodic disturbance of the water surface, then groups of surface gravity waves emerge and become longer and longer (see Vol. 1, Fig. 12.79). Those groups consisting of long waves reach the bank sooner than those consisting of shorter waves, thanks to their higher group velocities.

Quantitatively, we can understand most simply the modulation of weakly or not-at-all periodic, transient processes in quasi-monochromatic wave groups, that is wave trains of limited lengths and finite spectral widths, by considering a spectrometer which is based on a *line grating*. We thus want to treat the observation of a continuous spectrum of first order in such a spectrometer.

Figure 22.8 shows a grating with N lines or openings. A collimated light beam from an incandescent light source is incident on this grating, perpendicular to its surface. The line A in this case does not indicate a wave crest, but rather the limiting case of a non-periodic, transient pulse with the profile a as shown in Fig. 22.9. This type of process is shown especially clearly in Fig. 22.9; for that reason, we have chosen it as an example. A second pulse, which precedes the first, has already passed through the grating and has been split there into N pulses, each with the same profile. These pulses propagate in Fig. 22.8 in the form of eccentric circles to the lower right; however, only a segment of the circles is drawn in the figure. Along the direction of an arrow r (or v), the series of these pulses forms a wave group whose form is not sinusoidal, or more concisely, a non-sinusoidal wave group. It is sketched as curve b in Fig. 22.9, but only for $N = 6$. The spacing of the two pulses is large in the r direction and smaller in the v direction. Along the r direction, it could for example be 0.75 µm, and along the v direction only 0.4 µm.

Every such non-sinusoidal wave group b can be represented as a superposition of groups of sinusoidal waves of equal lengths (FOURIER decomposition), with the wavelengths λ, $\lambda/2$, $\lambda/3$, etc. This is shown in Parts c, d, e, f of the figure (compare Vol. 1, Sect. 11.3).

Now we come to an essential point: We want to observe visually, by *eye*; but the eye acts selectively, i.e. it chooses what it sees. It re-

[2] They can employ dispersion, absorption, or reflection in "filters".

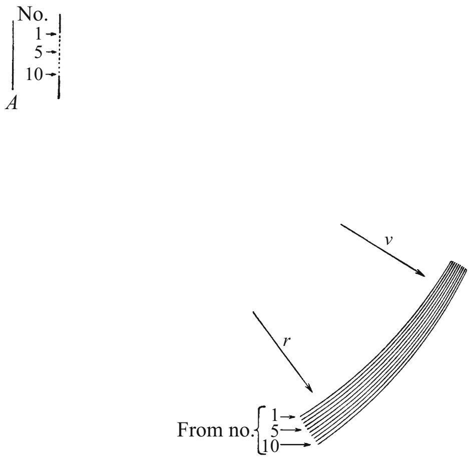


Figure 22.8 The production of wave groups by a grating. The line *A* indicates a non-periodic process (pulse) which is incident on the grating from the left, perpendicular to its surface, as can be demonstrated for example by a ground swell in shallow water or by an ultrasonic boom in the air. Its profile is sketched in Part *a* of Fig. 22.9, at the upper left.

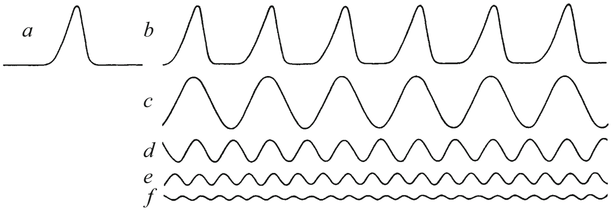


Figure 22.9 The periodic but not sinusoidal wave group *b* can be thought of as a series of 6 pulse-shaped groups *a* emitted by 6 openings or lines of a grating. In Parts *c* to *f*, the periodic group *b* is decomposed into four sinusoidal, monochromatic wave groups (FOURIER components).

sponds only to wavelengths between 750 nm and 400 nm. As a result, in the *r* direction, we see only one sinusoidal wave group (curve *c*) at $\lambda = 750\,\text{nm}$ (red), and in the v direction, at $\lambda = 400\,\text{nm}$ (violet). Thus, we can say concisely but unmistakeably that continuous spectra consist of groups of sinusoidal waves produced by the grating.

We often experience the corresponding acoustic experiment (TH. YOUNG, 1801; J.J. OPPEL, 1855) on the street. If we walk on a hard stone surface alongside a garden fence, we can hear with each step a whistling *chirp* of noticeable length. The fence acts as a reflection grating. Each slat reflects the air pulses from our footsteps, and thus the grating converts the non-periodic impulses into a non-sinusoidal wave group. Our ears are much less selective than our eyes; the ear responds to roughly 10 octaves. It thus responds not only to the longest wavelength λ, but also to $\lambda/2$, $\lambda/3$, etc. We therefore hear the non-sinusoidal wave groups as a *chirp*, and not as a tone or a note, which we would hear with a sinusoidal profile (Vol. 1, Sect. 12.28).

The number of individual wavelets (i.e. a wave crest + trough) in the group produced by the grating is, in the first-order spectrum, equal to the number of openings or lines N in the grating. N however, from Eq. (22.3), is equal to the resolving power $\lambda/\mathrm{d}\lambda$ in the first-order spectrum. Then the *resolving power* acquires a simple meaning: It is the *number of individual wavelets which are produced by the grating from a non-periodic process and can be combined into a group*. This holds not only for gratings, but also for spectrometers of all kinds. This statement can be verified with the help of all those interference experiments in which light from a *narrow* region out of a *continuous* spectrum is employed.

22.5 Comparison of Prisms and Gratings

According to the numerical example given at the conclusion of Sect. 22.1, with *one* prism of 10 cm base length, we can obtain a resolving power of $\lambda/\mathrm{d}\lambda \approx 10^4$. Connecting several prisms in series, e.g. 3, we can obtain three times that resolving power, that is around $3 \cdot 10^4$. With gratings, or in general with interference spectrometers, around ten times higher resolving powers are attainable, that is a resolving power of several 10^5.

In a comparison between gratings and prisms, however, we must not consider the resolving power alone. Another very important characteristic is the *usable wavelength range* $\Delta\lambda$. A prism always produces only a single spectrum. In it, each direction corresponds to only *one* wavelength. A grating, in contrast, produces a whole series of spectra of different orders m, and all of these spectra overlap. Each direction corresponds to several wavelengths, namely λ for $m = 1$, $\lambda/2$ for $m = 2$, $\lambda/3$ for $m = 3$, etc. A unique correspondence between wavelength and direction is obtained only within a certain limited range $\Delta\lambda$.

Let us look at Fig. 22.6. A spectral line of wavelength $(\lambda + \Delta\lambda)$ and order m must be spaced at least in the minimum β just adjacent to the spectral line of order $(m + 1)$ and wavelength λ; otherwise, the unique assignment of the spectral lines to deflection angles would be lost. We thus find

$$m(\lambda + \Delta\lambda) = (m + 1)\lambda - \frac{\lambda}{N},$$

or, when λ/N can be neglected relative to λ, for the usable wavelength range, we obtain:

$$\Delta\lambda = \frac{\lambda}{m}. \tag{22.4}$$

The most favorable case is found for $m = 1$; then, we find $\Delta\lambda = \lambda$. This means that a spectrum of *first* order has a unique assignment between wavelengths and angular deflection over a range from λ to 2λ, that is a full octave. If there are wavelengths outside this octave range, they must be sorted out individually in some manner.

For observations by *eye* (in contrast to photographic registration or an electronic detector), we need no special auxiliary apparatus for this sorting process. Our eyes themselves act selectively; they respond only to waves in the range of about one octave (ca. 400 to 750 nm). As a result, the eye can capture a whole first-order spectrum without hindrance.

The situation is however quite different for spectra of higher order, e.g. $m = 3$: Here, the usable wavelength range $\Delta\lambda$ is equal to only $\lambda/3$. Therefore, even the eye requires a *preselection* by some sort of additional apparatus, which must sort out the undesired waves. Often, a filter suffices for this task. For $m = 3$, it must transmit for example only waves between 450 and 600 nm, or between 600 and 800 nm, etc. (cf. Sects. 20.12 and 20.13).

22.6 Various Forms of Line Gratings

The line grating was developed in 1821 by J. FRAUNHOFER into the indispensable scientific instrument that it remains today. A FRAUNHOFER grating makes use of small orders m, usually between 1 and 5, and has a large number of grating openings or lines. Modern gratings have up to $N \approx 1.5 \cdot 10^5$. This allows us to achieve a resolving power of $3 \cdot 10^5$ (Eq. (22.3)) even in the second order. This means that the grating is able to separate two light signals with a wavelength difference of only about 3 millionths of their average wavelength. The usable wavelength range $\Delta\lambda$ remains rather large; in the second order, $\Delta\lambda = 0.5\ \lambda$. We can for example observe a spectrum in the visible light range from 750 to 400 nm in a single measurement. The structure of complex line spectra, such as those from atoms, are best investigated using large FRAUNHOFER gratings (see Fig. 22.4).

All of the lines or openings of the grating must be encompassed by the surface of a lens or concave mirror. Lenses and concave mirrors are only seldom available in the laboratory with diameters of more than 15 cm. For this (financial) reason alone, the openings of the FRAUNHOFER grating must be placed very close together, and all $1.5 \cdot 10^5$ openings must be next to each other on a surface area of ca. 15 cm diameter. This cannot be accomplished as in the construction of a garden fence using slats and spaces. Instead, the ruling of the grating is carried out in the form of parallel *grooves* on a highly-polished metal surface. This is achieved with a ruling engine using a diamond stylus. In this way, H.A. ROWLAND achieved 800 grooves per millimeter in 1882, an astonishing accomplishment for grooves 10 cm long! The ruled grating (e.g. with up to 1200 grooves per mm) is best employed as a *reflection grating*. Often, it is also used as a matrix for pressing *transparent* gratings in plastic. Gratings are often

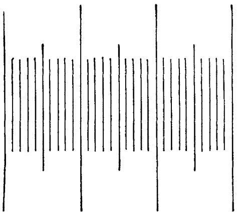


Figure 22.10 A piece of this rough millimeter scale, etched into glass, about 15 cm long is sufficient at grazing incidence of the light beam to cleanly separate the lines in a mercury spectrum (actual size)

ruled onto a concave metallic mirror; with such a "concave grating", one avoids the need for a lens in front of the grating.

For *X-rays*, the usual optical gratings made of glass or metal can be employed, as long as the wavelength λ is greater than 2 nm ($\hat{=}$ 620 eV).C22.2

They are used as reflection gratings at grazing incidence, since the ruling of the grating is fine enough for X-rays only if it is strongly foreshortened as seen in perspective at grazing incidence.

Perspective foreshortening of the ruling of a grating can be demonstrated by a simple experiment. We use the millimeter divisions on a common ruler as a diffraction grating for visible light (Fig. 22.10). At grazing incidence, the lines of a mercury spectrum can be cleanly separated.

In using mirrors and gratings for X-rays, another point must be kept in mind: The refractive indices of all materials are close to 1 for X-rays, and therefore their reflectivities are vanishingly small. But a favorable circumstance comes to our aid here: The refractive index of all materials for X-rays is somewhat *smaller* than 1 (Sect. 27.9). As a result, at nearly grazing incidence, we obtain *total reflection* of the X-rays.

We mention only one of the many possible variations on line gratings:

The *mirror surface grating* (blazed grating):

In a highly reflective metal surface, grooves are cut with a one-sided triangular profile ("blazing"), for example as shown in Fig. 22.11. A collimated light beam 1 is incident along the normal to the grating. The greater portion of its energy is reflected along direction 2, according to the law of specular reflection. With a suitable choice of the grating constant d, the first-order spectrum can be adjusted to fall around this direction. It then contains a much larger portion of the radiant power than the spectra of all the other orders on both sides of the normal to the grating; this grating has practically only *one* spectrum. Such mirror-surface or blazed gratings can be fabricated especially successfully for the long waves of infrared light (λ = ca. 10 to 300 μm), but they can also be made for the visible spectral range.

C22.2. The unit eV (electron volt) is a unit of *energy*. It is equal to the product of the magnitude of the elementary electric charge e_0 and the unit of potential difference, volt: 1 eV = $1.6 \cdot 10^{-19}$ A s · V = $1.6 \cdot 10^{-19}$ W s. Its relation to the wavelength is found from

$$E = h\nu = hc/\lambda$$

(h is PLANCK's constant, c is the vacuum velocity of light).

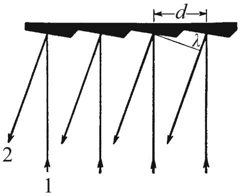


Figure 22.11 A mirror surface grating (also called a "blazed grating" or an "echelette")

22.7 Interference Spectrometers

Many spectral lines do not have the simple shape treated in Sect. 22.3. Instead, they consist of a number of components, which often overlap with each other. Frequently, a strong spectral line is also flanked by weaker “satellites”. Briefly put: Many spectral lines have a complex structure. The experimental investigation of *line structures* requires on the one hand the resolving power $\lambda/\mathrm{d}\lambda$ of a large grating (Eq. (22.3)), but on the other hand, a small usable wavelength range $\Delta\lambda$ suffices (Eq. (22.4)). Therefore, at low orders m, it is not necessary to make the number N of interfering wave trains very large. A smaller N and a higher m are sufficient; i.e. a large path difference $m\lambda$ between two neighboring wave trains. This is experimentally much simpler: We first limit the wavelength range under investigation by *preselection*. This means that we separate out the spectral region to be investigated from the rest of the spectrum, usually with a prism instrument; sometimes a filter is sufficient. The remaining light is passed through a thick, plane-parallel *air plate* (Fabry-Pérot étalon), as described in Sect. 20.10 and Fig. 20.15. By multiple reflections between two plates (with mirror surfaces), we generate a large number N of interfering light beams. The observations are carried out at nearly normal incidence ($\beta \approx 0$) with transmitted light, and yield bright spectral lines on a dark background (spectrometers of this design are named for Ch. Fabry and A. Pérot). The path difference of neighboring wave trains is, depending on the thickness of the air plate used, generally some ten-thousands of wavelengths (plate spacing several cm). This means that the spectral lines are formed via interferences of orders m between 10^4 and 10^5. Therefore, the usable wavelength range $\Delta\lambda = \lambda/m$ is less than $10^{-4}\lambda$.

A variation on this spectrometer design is named after Lummer and Gehrcke. The light leaves a plane-parallel glass plate under a large angle, i.e. at a grazing angle. Within the plate, it is reflected at nearly the critical angle for total reflection. This makes it possible to obtain multiple reflections without applying a reflective coating to the glass surfaces (Fig. 22.12).

Mirror images as grating arrangements of wave centers (sources) are of eminent importance for X-ray spectroscopy at wavelengths of less than ca. 2 nm. In Vol. 1, we showed how a grating pattern of mirror images could be produced by plane-parallel plates whose surfaces consist of lattice planes. We refer to Fig. 12.66 there, and repeat

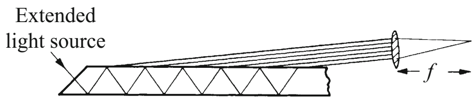


Figure 22.12 A Lummer-Gehrcke interference plate as a spectrometer (schematic). The rays of a collimated light beam are sketched.

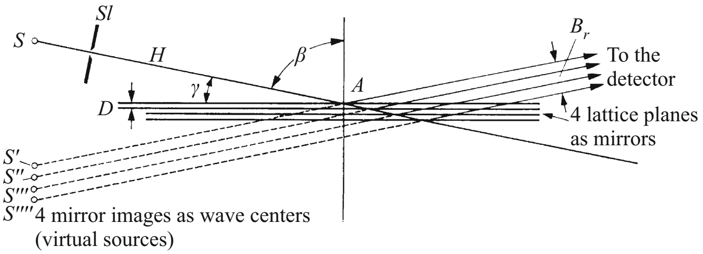

S
Sl
H
β
γ
A
D
B_r
To the detector
4 lattice planes as mirrors
S'
S''
S'''
S''''
4 mirror images as wave centers (virtual sources)

Figure 22.13 A Bragg spectrograph for X-rays. S is the line focus of an X-ray source which is perpendicular to the plane of the page (Sect. 19.5 and Fig. 19.10). H is a narrow beam of radiation, usually selected by a series of slits Sl, and not just its central ray. The receiver is a photographic plate or one of the radiation detectors mentioned in Sect. 15.4. The spacing D of two lattice mirror planes (only 4 are drawn here) is of the order of $3 \cdot 10^{-7}$ mm. As a result, the beams reflected from many lattice planes overlap. All together, they have an overall diameter B_r which is hardly greater than that of the incident beam. Therefore, even without imaging, we obtain sharp but very weak spectral lines. For a *demonstration with visible light*, the series of reflecting surfaces shown in Fig. 20.22 are suitable, or also a sequence of reflecting layers produced within photographic plates by standing waves (see Fig. 20.25).

the schematic here in Fig. 22.13 with a larger angle of incidence β. Lattice planes are provided by nature in crystals with a high degree of perfection. Their mirror planes occur with a large number N of layers, one above the other. This series of mirror planes yields a large number of mirror images which are arranged as a grating and can serve as wave centers (virtual sources).

For X-rays, *there are no lenses, and we therefore cannot use the* Fraunhofer *observation mode*. This means that we cannot separate broad light beams which have only small angles of inclination relative to each other in the focal plane of a lens (Sect. 16.8). As a result, we cannot readily use extended sources of X-rays together with plane-parallel plates. A narrow, line-shaped source must be employed, and the angles β, under which the individual wave trains are reflected from the same position on the crystal, must be scanned *successively* by rocking the crystal plate (Fig. 22.13). Generally, one quotes the grazing angle $\gamma = 90° - \beta$ instead of β itself. We find[3]:

$$\cos\beta = \sin\gamma = \frac{m\lambda}{2D}\,. \qquad (21.2)$$

[3] D is the spacing between two neighboring lattice planes (the optical grating constant, here also a lattice constant of the crystal); e.g. $D = 0.28$ nm in a NaCl crystal. The *crystallographic* lattice constant a, by contrast, is the spacing between two similar lattice units in homologous positions. In a NaCl crystal lattice, this is the spacing of two neighboring Na+ ions or Cl− ions. $a = 2D$ is 0.56 nm in the NaCl lattice. A cube of edge length a forms the *unit cell* of the NaCl lattice. This means that we can construct the entire crystal lattice simply by translating the unit cell parallel to the edges of the crystal.

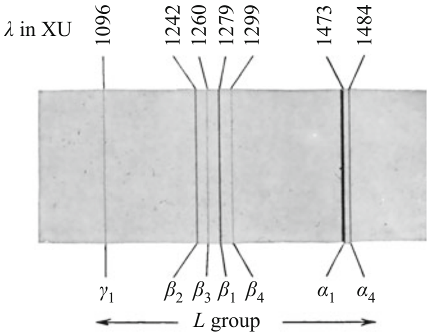


Figure 22.14 The line spectrum of L-shell radiation from tungsten, photographed with a vacuum spectrograph (using a calcite crystal with $D = 0.3029$ nm; photographic negative, 1 XU (X-ray unit) = $1.00302 \cdot 10^{-13}$ m)

For demonstration experiments, it is best to use a fluorescent crystal as detector, e.g. Tl-doped NaI. Its fluorescence radiation is measured with a sensitive photocell or a photomultiplier tube (Sect. 15.4). Figure 22.14 shows an example of a line spectrum in the X-ray region (See also the DEBYE-SCHERRER method, Sect. 21.14).

Exercises

22.1 The grating in Fig. 22.4 is assumed to have a grating constant of $D = 40\,\mu$m. On a screen 2 m away, the first-order maximum ($m = 1$) appears at a spacing of 3 cm from the central maximum ($m = 0$).
a) What is the wavelength λ of the light used?
b) After the whole apparatus is immersed into a liquid, the spacing of the first-order maximum is only 2.25 cm. What is the refractive index of the liquid?
(Sect. 22.2)

23 The Velocity of Light, and Light in Moving Frames of Reference

23.1 Preliminary Remark

In 1676, Olaf Römer, a Dane, at that time tutor to the royal princes at the court of Louis XIV in Paris, discovered the finite propagation velocity of light. Based on astronomical observations, he obtained a value which is of the correct order of magnitude, namely $c = 2.3 \cdot 10^8$ m/s. He used light signals from one of the moons of Jupiter at the moment when it emerged from the shadow of the planet. The delay between successive signals was found to be 42.5 h, i.e. equal to the time required for one orbital passage of the Jupiter moon. He carried out the measurements when Jupiter was at its closest approach to the earth and when it was furthest away, that is at opposite points along the earth's orbit around the sun, separated by the diameter of the orbit (the diameter of earth's orbit is $3 \cdot 10^{11}$ m). At the most distant point on the orbit, the signal showed an additional delay of 1320 s. From this, he was able to calculate the above value of the velocity of light.

Römer's achievement is worthy of admiration even today. We now know that light is an electromagnetic wave. Modern communications technology sends electromagnetic waves around the circumference of the earth, passing along a great circle of length $l = 40\,000$ km in a time $t = 0.133$ s. It follows that $c = l/t = 3 \cdot 10^8$ m/s.

Measurements today within the electromagnetic spectrum encompass a frequency range from around 10^{22} Hz (γ rays) down to about 10^5 Hz (long radio waves, employed by communications technology). According to various types of precision measurements, the current best value for the phase velocity of electromagnetic waves, and thus also for the velocity of light in vacuum,[C23.1] is $c = 2.998 \cdot 10^8$ m/s.

[C23.1] The currently accepted precise value is $c = 2.99792458 \cdot 10^8$ m/s.
In 1983, in the framework of the new definition of the unit of length, the meter, the velocity of light in vacuum was fixed at this value (see Comment C1.5 in Vol. 1). Fixing this value however does not mean that methods for determining c have become superfluous in textbooks; after all, the quoted value was found from the corresponding precision measurements. The fixed value means only that for the meter, there is currently no more precise definition available through other methods of measurement.
Another consequence is that due to the relation $c^2 = \varepsilon_0\mu_0$ (see Eq. (8.7)), and resulting from the fixing of the magnetic field constant μ_0, the value of the electric field constant ε_0 is also fixed (cf. Sects. 5.3 and 2.13).

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_23

23.2 Example of a Measurement of the Velocity of Light

The *phase velocity c of light in vacuum has in the course of time acquired a fundamental importance for all of the physical constants.* For this reason alone, a description of its measurement should not be lacking in an introductory course on physics. We describe here a method which was introduced by the physician L. FOUCAULT in 1850. In this method, a light beam passes along a path of known length in both directions and the corresponding transit time is measured directly using a uniform rotation of known rotational frequency. Figure 23.1 shows the experimental setup. It makes use of a telecentric light path with a clearly-defined position of the pupil.

S_0 is the light source, an illuminated slit. The axle of a small rotatable mirror is at the focal point of the lens L. With (initially slow) rotation, the mirror emits N light signals in a time t into the opening of the lens L. Each signal produces an image S' of the slit on the planar mirror P. Following reflection there, the light beam takes the same path in the reverse direction and at its end, it projects an image S'' of the first slit image S'. This second image S'' lies within the slit S_0, and is thus not visible. But with the aid of a half-silvered mirror H (a thin plane-parallel glass plate), the second image can be shifted to the side and projected onto a screen. Using this auxiliary mirror H, we can see the principle of the setup; for this purpose, the rotatable mirror is turned slowly by hand back and forth. The segment α of the light beam moves in the direction of the curved double arrow. At the same time, the segment β of the light beam is shifted parallel to the optical axis. Both movements are indicated for the beam segments α' and β'. The first image of the slit S' moves across the whole diameter of the planar mirror P in the direction of the straight double arrow. *In spite of these motions of the light beam and the first image S', the*

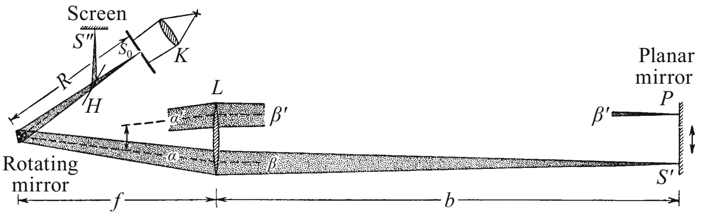


Figure 23.1 Measurement of the velocity of light by the method of FOUCAULT (1850), as simplified by A.A. MICHELSON (1878). The rotating mirror serves as the entrance pupil, and the light path to the right of L is telecentric. (A numerical example: $R = 5.2\,\text{m}$, $f = 10.5\,\text{m}$, $b = 32\,\text{m}$, diameter of L and P is 30 cm each, diameter of the rotating double-sided mirror 5 cm, rotational frequency of the mirror up to about 200/s. The rotational frequency N/t of the circulating light beam of radius R is up to 400/s. The displacement s of the image of the slit is up to ca. 4 mm).

second image S'' remains unchanged and at rest. This the decisive point. The reason is not difficult to understand: At low rotational frequencies, each light signal strikes the rotatable mirror on its return path at practically the same position as on its outward path.

This changes at high rotational frequencies. The returning signal finds that the mirror has rotated through a small angle during its transit time. Therefore, the image S'' of the slit is displaced by a small distance s to one side. If we now remove the auxiliary mirror H, we find S'' no longer directly on the slit S_0, but rather displaced to one side by s. We have:

$$\text{Transit time} = \frac{\text{Path traversed}}{\text{Velocity}}$$

or

$$\frac{t}{N} \cdot \frac{s}{2\pi R} = \frac{2(f+b)}{c}\,. \qquad (23.1)$$

The data that apply to the lecture hall in Göttingen can be found in the caption of Fig. 23.1.

23.3 The Group Velocity of Light

In Fig. 23.1, we can place part of the light path in a strongly dispersive liquid, e.g. carbon disulfide. Then for the nearly non-periodic wave groups of *incandescent light*, the same thing holds as for the similar groups of gravity waves on a water surface: They are ‘stretched out’ into long wave groups. Their front flank consists of longer, nearly sinusoidal waves, while their rear flank consists of shorter waves. The “red” light (refractive index $n \approx 1.6$) arrives first, while the “violet” light ($n \approx 1.7$) arrives last. As a result, in Fig. 23.1, the second image S'' of the slit appears as a short spectrum.

Only the *vacuum* is strictly *dispersion-free*. An experimental proof of this: When a Jupiter moon emerges from behind the planet, we see it immediately as colorless, not first red, then yellow, green, and so forth.

When dispersion occurs, we can no longer measure the phase velocity of waves simply in terms of path length and transit time. Instead of the phase velocity c, we obtain only their group velocity c^*. This concept, which is important for both physics and technology, was explained in detail in Sect. 12.22 of Vol. 1. A group velocity c^* can be defined for waves only within a limited spectral range. Within that range, we find:

$$c^* = c' - \lambda\frac{\mathrm{d}c'}{\mathrm{d}\lambda}\,; \qquad (23.2)$$

with $c' = c/n$, and (after differentiation)

$$\frac{\mathrm{d}c'}{\mathrm{d}\lambda} = -\frac{c}{n^2}\frac{\mathrm{d}n}{\mathrm{d}\lambda}\,. \tag{23.3}$$

(c is here the phase velocity in vacuum, and $c' = c/n$ is the phase velocity in a material of refractive index n).

Example: In carbon disulfide (CS_2), yellow light of wavelength $\lambda = 589\,\mathrm{nm}$, the well-known "sodium light" or "sodium D-light", has a refractive index of $n_\mathrm{D} = 1.63$. The phase velocity of this light is therefore $c'_\mathrm{D} = c/1.63 = 1.84{\cdot}10^8$ m/s. However, the measured value is only $1.72{\cdot}10^8$ m/s. This is the *group velocity* c^*_D of light in this wavelength range. In order to calculate it from Eq. (23.3), we must know the experimentally-determined dispersion of CS_2 for sodium D-light as well as the phase velocity c'_D. We find $(\mathrm{d}n/\mathrm{d}\lambda)_\mathrm{D} = -1.88\cdot10^5$/m and $c'_\mathrm{D} = c/n = 1.84\cdot10^8$ m/s. With these values, using Eq. (23.2) we obtain the result $c^*_\mathrm{D} = 1.72\cdot10^8$ m/s.

23.4 Light in Moving Frames of Reference

1. For a *light source outside the moving frame of reference* (astronomical aberration)

The earth's orbit around the sun is a large carousel, which appears pointlike when observed from a star. Figure 23.2 shows the earth at two arbitrary points along its orbit, separated by one-half of a year. From these points, an angle δ is observed between a star near the earth's *orbital axis* and a star near the earth's *orbital plane*. One observes (Fig. 23.3) an angular difference of $2\gamma = \delta_\mathrm{D} - \delta_\mathrm{J} = 41$ arc second, i.e. an *aberration* (the ratio $\gamma = u/c \approx 10^{-4}$). The velocity u_S of the sun relative to the system of fixed stars is unknown; only the difference $2u \approx 60\,\mathrm{km/s}$ is known.

As a result of these variations, all fixed stars near the earth's orbital *axis* appear to move on a *circular orbit* with a diameter of 41 arc second. Stars *between* the orbital axis and the orbital plane

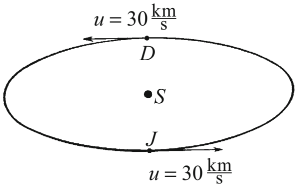

$u = 30\,\frac{\mathrm{km}}{\mathrm{s}}$
D
S
J
$u = 30\,\frac{\mathrm{km}}{\mathrm{s}}$

Figure 23.2 Variation of the velocity of the earth along its orbit around the sun, relative to a distant star (the orbit is seen in perspective from the side)

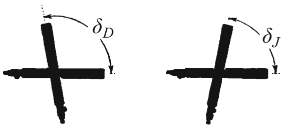
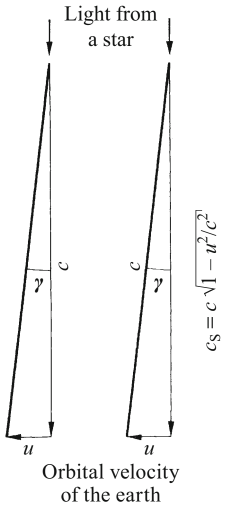


Figure 23.3 The apparent angular spacing δ between two stars changes according to the season of the year (*astronomical aberration*)C23.2

C23.2. Pointing the telescope: Think of raindrops which are hitting the window of a moving train. If we want to look at them directly, we have to fix our gaze along the direction of travel and upwards.

Figure 23.4 The measurement of the velocity of light using the aberration



follow elliptical orbits over the course of a year, with a major axis of 41 arc second. This phenomenon was discovered by BRADLEY and explained in 1728 as in Fig. 23.4, left side.

According to EINSTEIN's theory of relativity (1905; see Comment C5.1 in Chap. 5 of this volume), the velocity of light c is a limiting value which cannot be exceeded when adding velocities. Therefore, the left-hand diagram in Fig. 23.4 should be replaced by the diagram on the right. From it, we can read off

$$c_s = \sqrt{c^2 - u^2} = c\sqrt{1 - u^2/c^2} \tag{23.4}$$

and$^{C23.3}$

$$\tan\gamma = \frac{u}{c_s} = \frac{u}{c} \cdot \frac{1}{\sqrt{1 - u^2/c^2}} \,. \tag{23.5}$$

C23.3. Historical note: A derivation using the LORENTZ transformation (time dilation), was given in the 13th edition of POHL's "*Optik und Atomphysik*", Chap. 9, Sect. 4. English: see e.g. https://www.lsw.uni-heidelberg.de/users/mcamenzi/DopplerAberration.pdf or https://en.wikipedia.org/wiki/Stellar_aberration_(derivation_from_Lorentz_transformation) .

Now, however, $u \ll c$. Therefore, we can approximate the tangent by the sine and set the square root in the denominator equal to 1. Then the aberration becomes $\sin\gamma = u/c$ or $\gamma \approx u/c$. This result agrees with the observations of BRADLEY.

2. For a *light source* within *the moving frame of reference*

Figure 23.5 shows a carousel from above; at first, it is at rest. Two coherent light beams, 1 and 2, are emitted simultaneously from point A. They are reflected by mirrors at the vertices of the polygon to point B. There, they are superposed in a suitable manner, so that they produce interference fringes, for example curves of constant inclination. The position of the fringes is recorded (e.g. photographically). Then the carousel is set in motion, rotating in a counter-clockwise sense as seen from above, and the interference fringes are again photographed. Now, the fringes have shifted by some fraction Z of their original

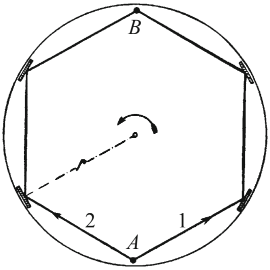


Figure 23.5 The measurement of the velocity of light by means of interference experiments on a carousel. In this simplified schematic, the components of the interferometer setup at points A and B are not shown.

spacing. From the magnitude of this shift, we can calculate the velocity of light.

Explanation: We choose our point of observation in the "laboratory frame", outside the carousel. Furthermore, we imagine that the light path along the edges of the polygon from A to B is replaced by the circumference of the semicircle, that is πr. Then we can say that each light beam has a transit time from A to B equal to $t = \pi r/c$. During this transit time, the target point B has moved forward at the velocity $u = \omega r$, and has covered the distance

$$s = \omega rt = \frac{\omega \pi r^2}{c} \tag{23.6}$$

($\omega = 2\pi N/t$ is the angular velocity of the carousel at its rotational frequency N/t. πr^2 is the area enclosed by the two light paths 1 and 2).

Therefore, light beam 1 must travel along a path which is longer by s, while light beam 2 travels along a path which is shorter by s. This gives rise to a path difference between the two light beams due to the rotation:

$$\Delta = 2s = \frac{2\omega \pi r^2}{c}\,. \tag{23.7}$$

This path difference in turn causes a shift in the interference fringes. It can easily be increased by a factor of 4. Firstly, points A and B can be placed adjacent to each other so that both light beams have to travel along the *full circumference of the carousel*. This doubles the shift of the fringes. Secondly, the sense of rotation can be reversed during the experiment, so that the shift of the fringes is again doubled. Then we find for the overall path difference

$$\Delta = \frac{8\omega \pi r^2}{c} \quad \text{or} \quad \frac{\Delta}{\lambda} = \frac{8\omega \pi r^2}{c\lambda}\,. \tag{23.8}$$

Numerical example: A path difference of $\Delta = \lambda/3$ is desired; with a reversal of the rotation, this will give a shift in the position of the fringes equal to 1/3 of their spacing. For yellow light, $\lambda = 0.6\,\mu\text{m} = 6 \cdot 10^{-7}$ m, and $c = 3 \cdot 10^8$ m/s. Then the product $N\pi r^2/t$ must be made equal to 1.2 m^2/s. This can be accomplished experimentally in various ways. Examples:

1. A carousel with an area of $1.2\,\mathrm{m}^2$ and $N/t = 1/\mathrm{s}$, that is one rotation per second.

2. The interference apparatus is put on board a ship. Its light path encloses an area of $\pi r^2 = 120\,\mathrm{m}^2$, and the ship sails around a full circle every 100 s, so that $N/t = 10^{-2}/\mathrm{s}$ (for rotational motion, the angular velocity is independent of the location of the axis of rotation; the latter is in the center of the apparatus in example 1., but off to the side for a circling ship).

3. The light path (protected by an evacuated pipeline) encloses an area of the order of $\pi r^2 = 10^5\,\mathrm{m}^2$. Then the angular velocity of the earth's rotation, $\omega = 2\pi N/t$, is sufficient; or, more strictly speaking, the angular velocity of its vertical component at the location of the observations. We thus have constructed an optical analog of FOUCAULT's pendulum demonstration (Vol. 1, Sect. 7.6).

Of course, we can neither bring the earth's rotation to a stop nor can we change its direction. Therefore, the determination of the original position of the interference fringes requires a trick: We first let the light beams follow a path enclosing a very small area, and then later use the large area. In this way, we lose a factor of 2 in Eq. (23.8), but nevertheless, the comparable experiment carried out by A.A. MICHELSON in 1925 yielded an impeccable result.

None of these methods is suitable as a demonstration experiment. Protecting the apparatus from perturbations by centrifugal forces and temperature variations requires a considerable effort. That is why we left it at the simple scheme described above, without considering the details of the light path.[C23.4]

[C23.4] The interferometer described here, named for G. SAGNAC, has considerable practical importance today as an "optical gyroscope" or "ring-laser gyroscope".

23.5 The DOPPLER Effect with Light

For *mechanical* waves, for example sound waves, either the source or the receiver or both may be moving relative to the *medium* of the waves, e.g. the air. Their velocity u can be clearly defined and measured. With all such motions, the frequency ν' measured at the receiver is different from the frequency ν emitted by the source. This is called the DOPPLER effect. When the receiver is moving and the source is stationary, we find (Vol. 1, Sect. 12.2):

$$\nu' = \nu\Big(1 \pm \frac{u}{c}\Big)\,, \tag{23.9}$$

where c is here the velocity of sound in the particular medium.

In contrast, when the source is moving and the receiver is stationary (the upper signs hold when the receiver is moving towards the source), we have:

$$\nu' = \frac{\nu}{\Big(1 \mp \frac{u}{c}\Big)} = \nu\Big(1 \pm \frac{u}{c} + \frac{u^2}{c^2} \pm \ldots\Big)\,. \tag{23.10}$$

C23.5. Herbert E. Ives and G.R. Stilwell, *Journal of the Optical Society of America* **28** (1938), p. 215. This experiment observes the *longitudinal* DOPPLER effect, i.e. parallel to the direction of wave propagation. There is also a *transverse* DOPPLER effect, but it is much smaller. Historical note: POHL discussed the transverse Doppler effect in the 13th ed. of "*Optik und Atomphysik*", Chap. 9, Sect. 6. For modern treatments in English, see e.g. mathpages.com/home/kmath587/kmath587.htm, or http://spiff.rit.edu/classes/phys314/lectures/doppler/doppler.html, or https://en.wikipedia.org/wiki/Relativistic_Doppler_effect#Transverse_Doppler_effect .

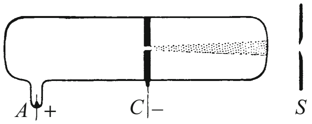


Figure 23.6 A simple ion-beam tube for observing the DOPPLER effect[C23.5] with light

We first consider only small values of the ratio u/c and therefore neglect the term u^2/c^2 and all the higher terms. Then Eqns. (23.9) and (23.10) are the same. The observed change in the frequency, $(\nu' - \nu)$, then depends only on the *relative* velocity u between the source and the receiver. This gives

$$\nu' = \nu\left(1 \pm \frac{u}{c}\right). \tag{23.11}$$

If we relax our restriction to small values of u/c, that is when our measurement precision also includes the second-order term u^2/c^2, then the equations (23.9) and (23.10) which hold for mechanical waves *cannot* be applied in optics. In optics, i.e. when dealing with electromagnetic waves, which are completely relativistic, we cannot distinguish between motion of the source and motion of the receiver, and there is no 'medium' that carries the waves. The two equations (23.9) and (23.10) must be replaced by a single equation; it is

$$\nu' = \nu\left(1 \pm \frac{u}{c}\right) \Big/ \sqrt{1 - \frac{u^2}{c^2}} = \nu\left(1 \pm \frac{u}{c} + \frac{1}{2}\frac{u^2}{c^2} \pm \ldots\right), \tag{23.12}$$

where c is now the velocity of light.

This equation can be derived from the LORENTZ transformations (see Chaps. 7 and 8 in this volume, in particular Comment C7.1 and Sect. 7.4, and the references quoted there). An experimental test of Eq. (23.12) was successfully carried out only in 1938, using the light emitted from ion beams.[C23.5]

Qualitative demonstration experiments can be performed using the ion-beam tube sketched in Fig. 23.6. The velocity u (of the ions in the beam) must be a few tenths of the velocity of light c.

Within the tube, and between the cathode C and the anode A, there is hydrogen gas at a low pressure of ca. 0.1 Pa. An anode voltage of around 30 kV produces a self-sustaining discharge in the gas. Accelerated, positively-charged hydrogen ions emerge from the channel in the cathode (canal rays). When these collide with the hydrogen molecules in the right-hand part of the tube, they produce rapidly-moving, excited hydrogen atoms which then emit light. The light is observed along the direction of motion of the ions with a spectrometer, leading to the image reproduced in Fig. 23.7. It shows a shift of the spectral lines towards shorter wavelengths, that is to higher frequencies.

This demonstration of the optical DOPPLER effect using mechanically-moving light sources does not justify the required efforts. It

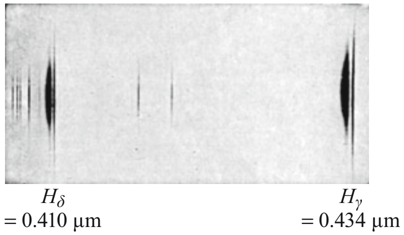


Figure 23.7 The DOPPLER effect in the spectrum of moving hydrogen atoms. The sharp lines H_δ and H_γ are emitted by atoms at rest (the BALMER series, see for example M. BORN's "Atomic Physics", 8th ed. (1969) (available for download – see https://archive.org/details/AtomicPhysics8th.ed). The broad lines which are seen to the left of the sharp lines are DOPPLER shifted from moving atoms and exhibit the distribution of their velocities.

shows no more than some arbitrary interference experiment in which a mirror is being moved at a velocity u along the direction of a light beam or opposite to it. As an example, we use an air wedge formed by two silvered plates (Fig. 20.11); we move one plate slowly and thereby continuously vary the path difference of the two wave trains that overlap between the plates. The interference fringes slide across the field of view, and the irradiation intensity at a particular *location* in the field of view varies periodically N times during the time t. Due to the DOPPLER effect, the frequency of the reflected wave train is shifted relative to the frequency ν of the incident wave train by $\Delta\nu = 2u/\lambda$. The superposition of the two wave trains thus produces beats of frequency $\nu_\text{S} = N/t = \Delta\nu$ (compare Vol. 1, Sect. 12.18).

The optical DOPPLER effect has acquired considerable importance for astronomy. In the spectra of distant stars and galaxies, the line spectra of known elements are often shifted towards longer or shorter wavelengths. This shift can be interpreted in many cases unambiguously as a DOPPLER effect. From its magnitude, the radial velocity u_r between the source and the earth can be computed. Generally large shifts, always towards longer wavelengths ("redshift"), are observed in the spectra of extragalactic objects (distant galaxies). They lead to surprisingly large radial velocities, up to several tenths of the velocity of light.C23.6

C23.6. Today, redshifts corresponding to much higher velocities have been observed. These are "cosmological redshifts" which are due to the expansion of space and not to mutual motion at a velocity *u within* space. They are a measure of the age of the emitting object, since the expansion has led with time to greater distances and greater velocities for distant objects. See e.g. http://curious.astro.cornell.edu/physics/104-the-universe/cosmology-and-the-big-bang/expansion-of-the-universe/610-what-is-the-difference-between-the-doppler-redshift-and-the-gravitational-or-cosmological-redshift-advanced or https://en.wikipedia.org/wiki/Redshift#Highest-redshifts .

All of these observed velocities are directed away from the earth, and their magnitudes increase *proportionally* to the distances of the galaxies. This is illustrated by Part a) of Fig. 23.8 (E is the earth). The lengths of the lines correspond to the velocities. Distances observable today are up to $5 \cdot 10^8$ light years.C23.7
This relation between distance and (expansion) velocity, discovered by E. HUBBLE, appears at first view to attribute an improbable special vantage point to our earth. But this is not the case; the graph a) could represent a race in which many students participate. Initially, they were all clustered around their teacher at position E. Then, at a given time, they all began to run away in all possible directions. Their goal is a distant circle with E

C23.7. The current record distance (2016) is about $3.3 \cdot 10^{10}$ light years, corresponding to a relative redshift of $z = \frac{\lambda' - \lambda}{\lambda} = 11.1$. Here, λ' is the observed wavelength and λ is the emitted wavelength (in the rest frame of the source). See the references in Comment C23.6.

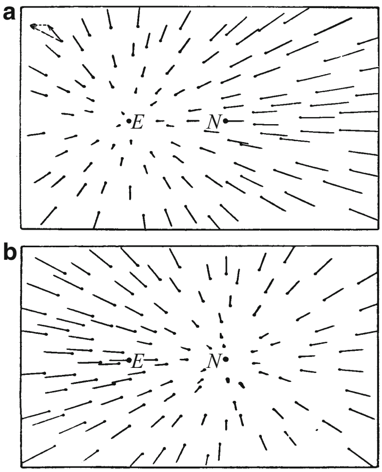


Figure 23.8 The radial escape motion of distant galaxies, obtained from the "redshifts" of their spectral lines (the location of the observer is at E in the upper graph, and at N in the lower graph)

at its center. At the moment of observation represented by the graph, each point in graph a) shows the position of one student, and his or her *velocity* is given by the length of the line. The distances covered since starting at E are proportional to the velocities of the runners. The fastest runners have moved furthest away. Part b) of the figure shows the same race, observed at the same moment in time, but now not from the location E of the teacher, but rather from that of some randomly-chosen participant in the race at the location N. The graph b) can be constructed very simply from graph a); we need only add the velocity vector of the runner at N in graph a) to all of the other velocity vectors in a) (shown at top left for one example as dashed lines). Now, E is no longer at the center point of the general radial escape velocities, but instead the location N is the apparent center point.

Polarized Light 24

24.1 Distinguishing Transverse and Longitudinal Waves

In the sections on mechanics, we learned to distinguish between transverse and longitudinal waves. Figure 24.1 shows as an example two "snapshots" or instantaneous images. The upper one represents a transverse wave, e.g. along an elastic cord; we see the crests and troughs of the waves. The lower snapshot shows a longitudinal wave, e.g. a sound wave in the air inside a pipe. We see compressions and rarifications[1]. A longitudinal wave exhibits the same behavior all around its direction of propagation (it is cylindrically symmetric relative to that direction), while a transverse wave, in contrast, is decidedly "one-sided". As shown in Fig. 24.1 (top part), it can be "linearly polarized". We want to see what this means in some detail.

View the wave perpendicular to its direction of propagation and consider the experiments. Initially, the viewing direction is also perpendicular to the plane of the page. Both wave phenomena appear in full clarity. Then, imagine that your viewing direction is *within* the plane of the page (from above or below the waves). The longitudinal wave still looks the same, but the transverse wave is now no longer visible; we see the cord as a straight line. The transverse wave in Fig. 24.1 thus has a one-sidedness or planarity which is called its "polarization". It is characterized by a single *plane of oscillation*. The motion of the transverse wave becomes invisible when the eye of the observer is itself in this plane of oscillation.

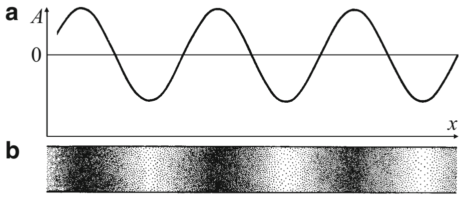
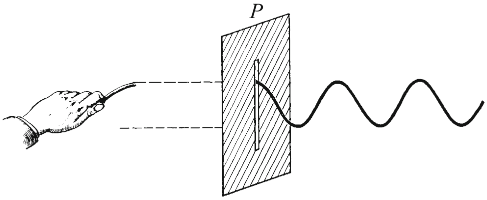


Figure 24.1 Snapshot a: A transverse wave (*A* is its deflection). Snapshot b: A longitudinal wave

[1] Figure 24.1 should be thought of as a "snapshot" of *two experiments*. In a drawing, any longitudinal wave can also be represented by a wavy line, e.g. a sine function. In the drawing of a sound wave, the ordinate may then represent the air pressure, that is wave crests refer to regions of compression; or it can represent the density, which is maximal at a wave crest.

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_24

Figure 24.2 A slit *P* as a polarizer for mechanical transverse waves



In mechanics, a polarization can thus occur only for transverse waves. But beware of the converse of this sentence: the lack of polarization does not rule out transverse waves. The position of the plane of oscillation of transverse waves can change rapidly and randomly. Then, *averaged over time*, the transverse waves may also have no polarization (we then call them "unpolarized").

Nevertheless, even in this case, we can distinguish experimentally between longitudinal and transverse waves. We make this clear again using a mechanical demonstration. In Fig. 24.2, a hand produces a transverse wave on a long elastic cord. The hand moves with a fixed frequency and amplitude, but it changes its direction of oscillation continually and randomly. As a result, the plane of oscillation of the waves also changes continually and randomly; the waves completely fill a cylindrical region with the direction of propagation as its center axis. The intersection of this cylinder with the plane of the page is indicated by two dashed lines. Now comes the essential point: At *P*, the cord passes through a narrow slit. This slit acts as a "polarizer". It selects one single fixed plane out of the mixture of rapidly changing planes of oscillation. In Fig. 24.2, this selected plane lies parallel to the plane of the page. Therefore, to the right of the polarizer *P*, we can observe a linearly-polarized wave. Its polarization clearly shows the character of the waves which are travelling from the left towards the polarizer: They are transverse waves.

24.2 Light as a Transverse Wave

The knowledge of waves that we obtained from mechanics can be applied analogously to optics. But – should we describe *light* in terms of longitudinal waves or of transverse waves?

We employ one of the fundamental observations of optics, the visible trace of a light beam in a cloudy medium. We can use water with fine suspended particles as such a medium. The light beam looks just the same all around its direction of propagation; we initially observe *no* polarization. But only a *positive* observation, i.e. the occurrence of a polarization, could eliminate longitudinal waves and demonstrate uniquely that light consists of transverse waves. We can obtain this positive evidence in the following way:

Erasmus Bartholinus, a Dane, discovered *birefringence* (also called double refraction) in 1669. He let a light beam *n* fall perpen-

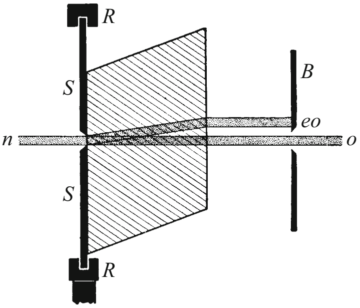


Figure 24.3 A demonstration of birefringence. A thick plate of calcite crystal (a natural rhombohedral cleaved crystal) is attached to a disk *SS* (seen here in cross-section). This disk can be rotated within the ring *RR* around the *n*-*o* direction as axis. When we add a circular aperture *B*, we have made a simple polarizer. (The direction of the optical axis of the calcite crystal is indicated by shading, as in Sect. 24.4)

dicularly onto a platelet of Icelandic calcite ($CaCO_3$ – Fig. 24.3). He observed that the beam was split into two sub-beams. One of the two, denoted by *o*, passes through the crystal platelet in its original direction without any refraction. It thus shows the same behavior as seen for any glass plate with a beam of light at perpendicular incidence. This sub-beam *o* is therefore called the "ordinary" ray. The other sub-beam *eo* experiences a refraction on entering the crystal, in spite of its perpendicular incidence, and it leaves the calcite crystal with a parallel shift of its beam axis. This second sub-beam is called the "extraordinary" ray.

There are several possibilities for eliminating one of the sub-beams. In the simplest case, the aperture *B* in Fig. 24.3 is sufficient. It allows only the ordinary ray (beam) to pass through. By eliminating one of the sub-beams, we have converted the doubly-refracting crystal into a *polarizer*. It serves the same purpose for light as the slit in Fig. 24.2 for the mechanical waves on the cord. We will see this in our next experiment. We allow the light to pass through such a polarizer and then follow its trace in a container filled with cloudy water (Fig. 24.4). Now, the light beam shows a clear-cut polarization: We can observe the light beam from a direction perpendicular to its propagation direction and look at it from all sides. Along two particular directions, separated by 180°, the beam becomes *invisible*; in these directions, the eye is in the plane of oscillation of the polarized beam. We mark the position of this plane of oscillation on our polarizer with a pointer ***E***. Now, we can make the observations more straightforward. We maintain our viewpoint and make use of the pointer to position the polarizer around the axis of the light beam as optical axis. This allows us to demonstrate the transitions between good visibility and complete invisibility of the light beam to a large audience in an impressive manner.

C24.1. The light beam is made visible by *scattering* (Chap. 26). JOHN TYNDALL, 1820–1893, an Irish physicist, observed that small suspended particles in liquids can produce light scattering (TYNDALL effect); see *Proceedings of the Royal Society* (London), Vol. 17 (1868/69), p. 223, an article which is still quite readable today.

Video 16.1:
"Polarized Light"
http://tiny.cc/5dggoy

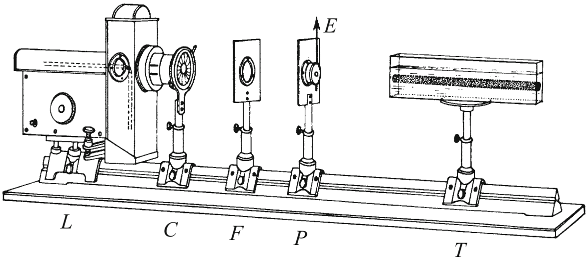


Figure 24.4 A demonstration of the plane of oscillation of a light beam. *P* is the polarizer. (The water is clouded by adding suspended particles, most expediently by using Styrofan (BASF), i.e. plastic spherules whose diameter is less than the wavelength of the light)C24.1 **(Video 16.1)**

We summarize: Using a polarizer, we can prepare light beams as transverse waves with a *fixed plane of oscillation*. The *light beam becomes invisible when the eye of the observer lies in its plane of oscillation*. This allows us to fix the position of the plane of oscillation in the polarizer and to mark it with a pointer.

The discovery of polarization considerably enriched the interpretation of light as a form of waves. We can now say that the wave scheme which we have often utilized, i.e. a wavy line, in the simplest case a sinusoidal curve, corresponds in optics to the picture of a *transverse wave*. Its "deflection" can be oriented parallel to a plane, i.e. the light wave can be linearly polarized. Therefore, the "deflection" and its maximum value, the "amplitude"[2] are directed quantities, vectors oriented transverse to the direction of propagation of the wave. We will correspondingly refer to the "deflection" of a light wave from now on as the *light vector* and denote it by the symbol $\boldsymbol{E}$.C24.2 Concerning the physical nature of the light vector, we need make no statements for the time being. We will continue to limit our descriptions of optical phenomena to what is absolutely essential.

C24.2. The letter $\boldsymbol{E}$ stands at the same time for the vector of the electric field with which we can describe light as an electromagnetic wave. See Sect. 12.6.

24.3 Polarizers of Various Types

The polarizer sketched in Fig. 24.3 produces light beams of only a few millimeter in diameter; for larger beams, we would need thick and expensive plates of calcite or some other double-refracting crystal. To overcome this disadvantage, a series of other types of polarizers have been developed.

In the first group, one or the other sub-beam is eliminated by *reflection*, making use of total reflection. For this purpose, we cut a piece

[2] Compare Sect. 16.9, conclusions.

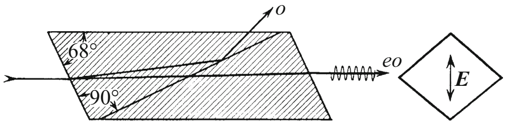


Figure 24.5 A NICOL prism, that is a polarizer after WILLIAM NICOL (1828), shown as a longitudinal section and a cross-section. A prism of this form is suitable for modest requirements. The extraordinary light beam is transmitted. Its plane of oscillation (the light vector ***E***) lies parallel to the shorter diagonal of the diamond-shaped cross-section. (The optical axis of the crystal is indicated by the shading)

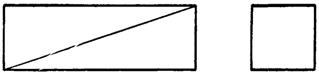


Figure 24.6 A polarizer with its end surfaces perpendicular to its long axis. In a superior construction as described by GLAN-THOMPSON, the quite uniformly-polarized field of view encompasses around 30°. Various but superficially similar shapes differ in the orientation of the crystal axes of the calcite. Therefore, one must determine the position of the plane of oscillation experimentally, e.g. as shown in Fig. 24.4, if the particular construction type is unknown.

of calcite crystal[3] in an oblique direction (Fig. 24.5) and separate the two halves by a transparent intermediate layer of suitable refractive index, i.e. one which leads to a total reflection of the ordinary rays (e.g. Canada balsam or linseed oil). For optimal performance, the two end surfaces should be perpendicular to the long axis (Fig. 24.6). With this shape, the transmitted sub-beam experiences no transverse displacement, so that when the polarizer is rotated, it does not "wobble".

In the second group of polarizers, one of the two sub-beams is eliminated by *absorption*. For this type, one makes use of "dichroic" materials. These are double-refracting and also absorb the two polarized sub-beams differently. In the most favorable case, one of the oscillating components is transmitted practically without attenuation over the entire visible spectral range, while the other, perpendicular to the first, is completely absorbed (see also Sect. 12.8). The most serviceable types of fabrication in use today are "polarizing foils" or "Polaroid sheets".[C24.3] One type of these foils contains many parallel-oriented tiny dichroitic crystallites. Another type consists of films of a transparent plastic with rod-shaped structural elements (miscellanea). These rods are oriented parallel to a particular axis during the manufacture of the foils by mechanical deformation and are treated with dyes which are absorbed onto their surfaces. With

C24.3. These polarizing foils are commercially available with dimensions of up to 1/2 m and more. They are quite suitable for demonstration experiments, e.g. using an overhead projector.

[3] All polarizers made from calcite are useless in the ultraviolet spectral range. Calcite, and especially the glue for attaching the parts, absorb strongly in the short-wavelength region. An alternative is shown in Fig. 24.9. In the infrared, calcite can be used out to $\lambda = 2.5\,\mu\text{m}$.

this process (E. KÄSEMANN), one can also fabricate polarization foils for the ultraviolet and the infrared spectral regions.[C24.4]

C24.4. Another type of Polaroid foil was invented in 1938 by EDWIN H. LAND. It is called "H-sheet" and contains a polyvinyl alcohol (PVA) polymer impregnated with iodine. During manufacture, the PVA polymer chains are stretched to form an array of aligned, linear molecules in the material. The iodine dopant attaches to the PVA molecules and makes them conducting along the length of the chains. Light polarized parallel to the chains is absorbed, and light polarized perpendicular to the chains is transmitted.

A third group of polarizers will be described in Sect. 25.6 (polarization by reflection).

24.4 Birefringence, in Particular in Calcite and Quartz

Polarized light plays a significant role in optics. We will meet up with it again and again in later chapters. Important accessories for producing and investigating polarized light are based on the phenomenon of *birefringence* in crystals (also called *double refraction*). It is thus expedient for us to consider some additional facts about the subject of birefringence.

Quartz crystals are generally known in the form of hexagonal columns. Calcite is also found in the same form, although its rhombohedral cleavage fragments are better known. We put two surfaces perpendicular to the long axis of the column and allow a thin light beam to enter the crystal parallel to its long axis. Then the beam passes through the crystal without deflection, the splitting into two spatially separated sub-beams is absent (Fig. 24.3). The long axis of the hexagonal column is thus optically distinguished; along this axis, there is no birefringence.

This special *direction* is referred to – not very adroitly – as the *optical axis*. ("Axis" here means a direction, not a line, deviating somewhat from the usual parlance!). Every plane which contains the optical axis is called a *principal plane* (or principal *section*) of the crystal[4]. We will often make use of this concept.

For our next experiment, we use two geometrically-identical calcite prisms as shown in Fig. 24.7. Within the upper prism, its optical axis lies parallel to the base plane of the prism; in the lower, it is perpendicular. This is indicated by the shading lines in the figure.

The light is normally incident on both prisms (perpendicular to the first surface, from the left). In the upper prism, it propagates parallel to the optical axis, and in the lower prism, it is perpendicular. As a result, birefringence occurs only in the lower prism, and only there do we see two separate beams. The beam that is more strongly deflected (o) takes a similar path as in the upper prism, without birefringence. Thus, it is the ordinary beam. The extraordinary beam (eo) is less strongly deflected. Both beams then pass through a polarizer P. Its direction of oscillation is marked by the double pointer $\boldsymbol{E}$. In the position shown, the polarizer allows only the extraordinary beam to pass. If we rotate it by 90° (so that $\boldsymbol{E}$ is perpendicular to the plane of

[4] In contrast to the principal plane of a prism, which is a plane perpendicular to the refractive edge of the prism (Sect. 16.6, beginning).

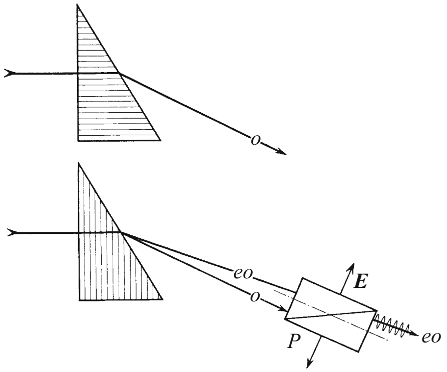


Figure 24.7 The birefringence of calcite. The direction of the "optical axis" is indicated by the shading lines, and the plane of oscillation of the extraordinary beam is shown pictorially. The principal plane of the prisms (a plane perpendicular to its refracting edge) is at the same time a principal plane of the crystals, here the plane of the page.

the page), then only the ordinary beam can pass through the polarizer. Therefore, the planes of oscillation of the two beams are perpendicular to each other. The plane of oscillation of the extraordinary beam lies along a principal plane of the crystal, while that of the ordinary beam is perpendicular to it.

From the angles of deflection, the refractive indices can be calculated. For green light, we obtain

$$n_{\mathrm{eo}} = 1.49\,,$$
$$n_{\mathrm{o}} = 1.66\,.$$

The extraordinary beam is less strongly refracted (Fig. 24.7, bottom). For this reason, calcite is called *negative* birefringent. For quartz, the opposite is the case; quartz is *positive* birefringent.

In Fig. 24.7, the beam in the interior of the crystal is either parallel to the optical axis (above), or it is perpendicular (below); i.e. the angle γ between the beam and the optical axis is either zero or 90°. The measurements can however also be repeated for intermediate values of γ, for example as shown in Fig. 24.8, left side. The refractive index n_{o} of the ordinary beam is found to be equal to the value quoted above, $n_{\mathrm{o}} = 1.66$, for all values of γ. The refractive index of the extraordinary beam, in contrast, changes with α and γ. It has its smallest value at $\gamma = 90°$, and its maximum at $\gamma = 0°$. For $\gamma = 0°$, $n_{\mathrm{eo}} = n_{\mathrm{o}}$, i.e. along the optical axis, birefringence is absent.

In Fig. 24.8, right side, a prism with a different orientation is sketched. In this case, the optical axis is parallel to the refractive edge of the prism, and thus perpendicular to the plane of the page. This is indicated by the dots. The two beams are perpendicular to the optical axis within the crystal for every angle of incidence α, i.e. γ is

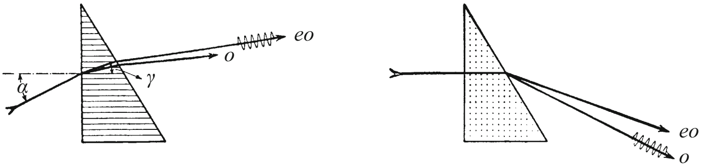


Figure 24.8 The birefringence of calcite. At left, the refractive indices can be measured for different angles of inclination γ between the beam and the optical axis. At the right, in contrast, γ is constant at 90°, because the optical axis lies parallel to the refracting edge of the prism.

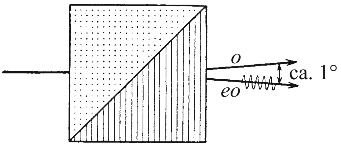


Figure 24.9 A double prism made of quartz gives two non-achromatized (see Sect. 18.8), symmetrically deflected sub-beams (WOLLASTON prism). When joined by a film of water, it is suitable for the polarization of ultraviolet light. For other choices of the axis directions in the halves of the prism, the ordinary beam can be made to pass through the prism without deflection and can thus be achromatized. However, this costs half of the beam divergence (ROCHON, SENARMONT prism).

always 90°. Therefore, for every angle of incidence, we measure the same two values of the refractive indices as given above, $n_\mathrm{o} = 1.66$ and $n_\mathrm{eo} = 1.49$.

The examples given thus far in this section refer to several special cases, which are also important for applications (Fig. 24.9): Both the first surface on which the light is incident as well as the plane of the page were either parallel or perpendicular to the optical axis. Without this condition, the situation becomes very complex even for single-axis ("uniaxial") crystals.

The essential point can be demonstrated as shown in Fig. 24.10. We make use of the same experimental arrangement as in Fig. 24.3, but the light is incident at an *oblique* angle and, together with the angle of incidence α, it determines a plane of incidence. In the orientation as drawn, the plane of incidence is parallel to a principal plane of the crystal. Both sub-beams propagate in the plane of incidence.

Now, the thick calcite block is rotated slowly around the normal *NN*. This causes the optical axis to move out of the plane of incidence. This is unimportant for the ordinary beam; it remains as before within the plane of incidence (plane of the page) over its entire path. However, the extraordinary beam continues to propagate along a principal plane of the crystal. This principal plane contains the normal and the optical axis. It is thus no longer in the plane of incidence and circles during the rotation around the ordinary beam on a cone inside and on a cylindrical surface outside. Apart from the special cases treated

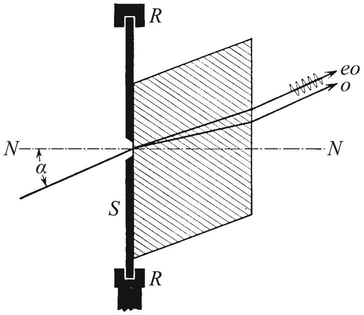


Figure 24.10 Refraction outside the plane of incidence. When the calcite block is rotated around the normal *NN* to the surface of incidence, the extraordinary beam is refracted out of the plane of incidence (the plane of the page).

above, the refraction of the extraordinary beam thus does not take place within the plane of incidence. The *elementary law of refraction* (Fig. 16.4) *fails*. The refraction of the extraordinary beam can in general be described only in perspective with a three-dimensional representation.

The phenomena become still more complex when the crystals have two axes, i.e. with crystals which have two internal directions that exhibit no birefringence. In such “biaxial” crystals, there is no “ordinary” beam at all. Both beams are “extraordinary”, i.e. for both, the refractive index depends on the direction, and both in general leave the plane of incidence when refracted. The planes of oscillation of both beams remain perpendicular to each other in biaxial crystals. For physical investigations, one often uses cleaved pieces of crystals from the biaxial group of clear *mica*[5].

Mica sheets have mechanically distinguished directions. If the sheet is laid onto a blotting pad and a hole is punched through it with a pin, a *chatter mark* as photographed in Fig. 24.11 results. It shows a six-pointed star with two long arms. The direction of the long arms is called the β direction, and the direction perpendicular to it in the plane is the γ direction.

The two beams resulting from birefringence oscillate parallel to the β direction and to the γ direction. Red-filter light which oscillates parallel to the β direction (which propagates faster within the crystal) has a refractive index of

$$n_\beta = 1.5908\,.$$

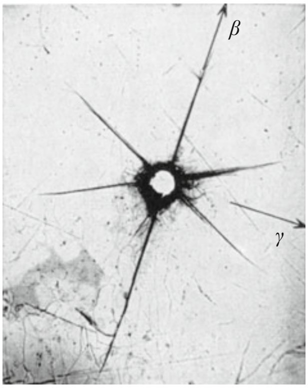
The red light which oscillates parallel to the γ direction (and propagates more slowly within the crystal) has a refractive index of

$$n_\gamma = 1.5950\,.$$

Many additional details of birefringence are important for crystallography, but not for its physics applications.

[5] The two optical axes in mica form an angle of 45° within the crystal. The midline of this angle is nearly perpendicular to the cleavage planes (deviations of less than 2°). The plane defined by the two optical axes intersects the cleavage planes in Fig. 24.11 along the direction γ.

Figure 24.11 A chatter mark on a sheet of mica



24.5 Elliptically-Polarized Light

In the sections on mechanics, we treated the superposition of two *perpendicular sinusoidal oscillations* (Vol. 1, Sect. 4.4; see also Sect. 9.4 and Fig. 9.20). When both oscillations have the same frequency, in general the deflection follows elliptical orbits; circles and lines are limiting cases. The form of the ellipses can be varied at will. We mention two methods:

1. The two mutually perpendicular sinusoidal oscillations x and y have amplitudes A and B; the phase difference δ is varied. In this case (Fig. 24.12), the axes of the ellipse lie at some angle between the directions of the two individual oscillations:

$$x = A\sin(\omega t + \delta)\,,$$
$$y = B\sin(\omega t)$$

($\omega = 2\pi\nu$ is the circular frequency).

Figure 24.13 shows examples for the special case in which $A = B$.

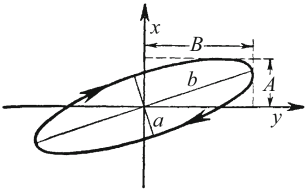


Figure 24.12 The formation of an elliptical oscillation by superposition of two mutually-perpendicular linear oscillations with the amplitudes A and B and a phase difference of $\delta = 45°$. The vertical oscillation along the x axis leads the horizontal oscillation. a and b are the semi-axes of the ellipse **(Video 9.1)**.

Video 9.1:
"Circular Vibrations"
http://tiny.cc/qcggoy.

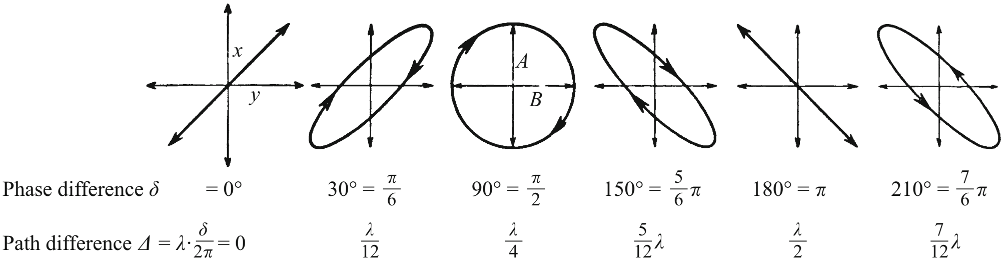
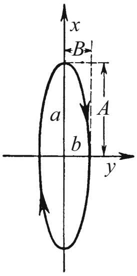


Figure 24.13 Examples of elliptical oscillations for the special case of $A = B$. The vertical oscillation *leads* the horizontal one by a phase difference of δ. That is, the vertical deflections begin sooner with positive values than the horizontal ones. If we wish to apply these images to travelling transverse waves (using the values of the path differences shown), they show the sense of rotation for light that is incident on the plane of the page in the positive z direction (that is perpendicular from above). Compare Figs. 24.15d and e.

Figure 24.14 The formation of an elliptical oscillation from two mutually-perpendicular linear oscillations with the amplitudes A and B and a phase difference of $\delta = 90°$. The semi-axes of the ellipse, a and b, are equal to the amplitudes of the linear oscillations, A and B.



2. The phase difference δ of the two individual oscillations is kept constant at 90°, and the ratio of their amplitudes is varied. Then the axes of the ellipses are parallel to the directions of the two individual oscillations (Fig. 24.14).

In a corresponding manner, we can superpose two propagating, linear polarized *waves*. Their planes of oscillation are positioned perpendicular to each other and their "light vectors" are added at every point along their path.

We will illustrate the composition of the waves and the forms of circular and elliptically-polarized waves with two examples using perspective drawings (Fig. 24.15). These represent *snapshots* – like all pictures of propagating waves. The direction of propagation is the z axis, from the front left to the rear right in the drawings (Fig. 24.15a).

In Fig. 24.15 b, the two waves have the same amplitudes and their path difference Δ is zero. When their vectors are added, we again obtain a linearly-polarized wave. Its plane of oscillation is inclined to the vertical by 45° (Fig. 24.15c). In Fig. 24.15 d, the two waves likewise have equal amplitudes, but now, the horizontally-oscillating wave leads the vertically-oscillating wave by $\Delta = \lambda/4$. Adding their vectors yields a *circularly-polarized* wave. In its snapshot image, the set of all the vectors makes up a helical surface or "spiral stairway"

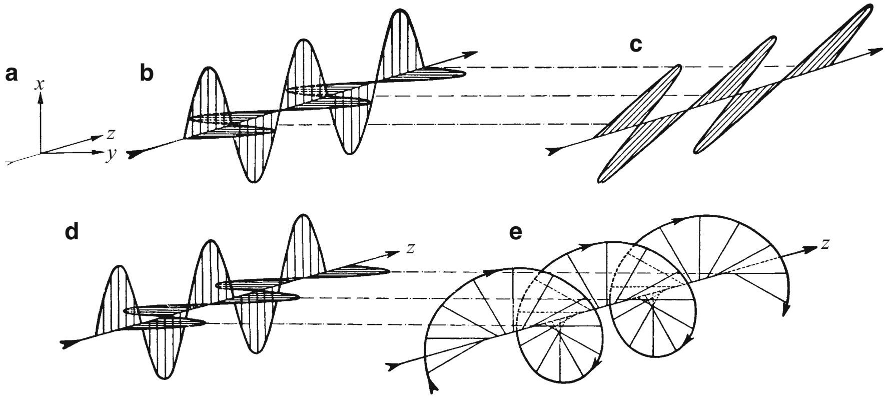


Figure 24.15 The superposition of two transverse waves which are oscillating in mutually perpendicular planes and propagating in the z direction. They have equal amplitudes (here "snapshots"; the positive x axis points upwards and the positive y axis horizontally to the right). In Part d, the horizontally-oscillating transverse wave leads the vertically-oscillating wave by $\lambda/4$. The arrowheads along the helix in Part e simply indicate its helical form. The *area of the helix does not rotate around the z axis as the wave moves forward. Instead, we should imagine the entire area enclosed by the helix to move in the z direction without rotation, at the velocity characteristic of the waves*. A fixed plane at the right rear which is perpendicular to z is then penetrated in sequence by the individual vectors (like the steps of a spiral stairway). Their lines of intersection rotate in a clockwise sense as seen by an observer looking *against* the propagation direction z (i.e. *towards the source of the waves*). This is a *right-circularly polarized* (rcp) light wave.[C24.5] If the horizontally-oscillating transverse wave were leading with a path difference of $3\lambda/4$, the snapshot would show a *left-hand* screw (a *left-circularly polarized* (lcp) light wave).

with the direction of propagation z as its center axis. In every pair of points which are spaced at a distance of one wavelength, the vectors point in the same direction; one rotation of the helix corresponds to one wavelength.

This general scheme, valid for every type of transverse waves, can be applied to the description of some important phenomena which are connected with birefringence. We demonstrate this with reference to Fig. 24.16. From the condenser C, a nearly perfectly collimated beam of light passes through a red filter F and onto a polarizer P. Its plane of oscillation, indicated by the pointer $\boldsymbol{E}$, is inclined by 45° to the vertical. The linearly-polarized light then strikes a birefringent mica platelet G at perpendicular incidence. In the mica plate, the light beam is split through its birefringence into two sub-beams. The beam which propagates faster within the crystal has its plane of oscillation vertical, while the slower beam oscillates in the horizontal plane. The two beams overlap almost completely within the small thickness d of the platelet, in contrast to Fig. 24.3, both inside the crystal and to the right, after leaving it.

After leaving the birefringent platelet G, the two light beams have a path difference (i.e. a difference in their optical path lengths; see

C24.5. The notation described in the figure caption is the so-called "optical convention"; the rotation of the electric field of the light is observed looking *towards the source*. The "physical convention" is just the opposite: There, the observer is supposed to be looking *away from the source* along the beam of light rather than *towards it*, so that the meanings of 'rcp' and 'lcp' are exchanged. Helpful animations can be seen at https://en.wikipedia.org/wiki/Circular_polarization
In this book, the *optical convention* is used.

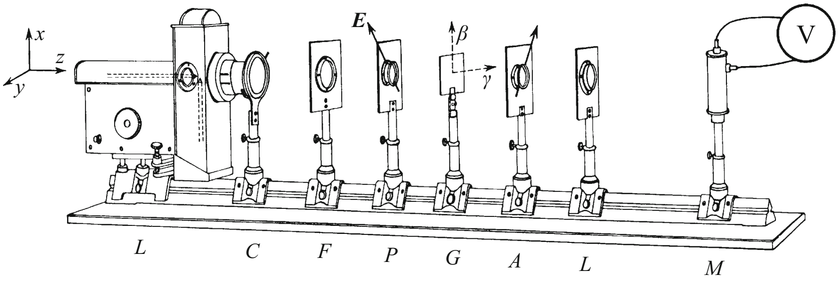


Figure 24.16 The production of elliptically-polarized light using a mica platelet G. β and γ are the directions defined in Fig. 24.11. Without the radiometer M, the arrangement is also suitable for demonstrating interference phenomena with collimated beams of light (Sect. 24.6).

Sect. 16.3) of

$$\varDelta = d(n_\gamma - n_\beta)\,. \tag{24.1}$$

We insert the values of the refractive index for red-filter light as given at the end of the previous section ($\lambda = 650\,\text{nm} = 6.5 \cdot 10^{-4}$ mm) and obtain

$$\varDelta = 42 \cdot 10^{-4} d$$

or

$$\frac{\varDelta}{\lambda} = \frac{42 \cdot 10^{-4}}{6.5 \cdot 10^{-4}\,\text{mm}} \cdot d = 6.5\,\frac{d}{\text{mm}}\,. \tag{24.2}$$

As a result of this path difference, the two light beams, which are oscillating perpendicular to each other, superpose to give an elliptically-polarized light beam (this includes of course the limiting cases of circularly- and linearly-polarized beams).

To identify the type of polarization, the section of the apparatus to the right of G is employed; its most important component is a second polarizer A, called in this application the “analyzer”. The light which it transmits falls on a lens L, and the lens forms an image of G either on the radiometer M (e.g. a photocell) or on an observation screen. So much for the experimental arrangement, now to its demonstration:

We set the analyzer into a slow, uniform rotation. At the same time, we observe the deflections of the radiometer for different angles ψ between the plane of oscillation of the analyzer and that of the polarizer. Examples:

1. A ‘dry run’ experiment without the mica platelet G (i.e. for $d = 0$). Only linearly-polarized light reaches the analyzer. It allows only the component of the light vector $\boldsymbol{E}$ of the incident light which is parallel to its transmission direction to pass through, with a magnitude

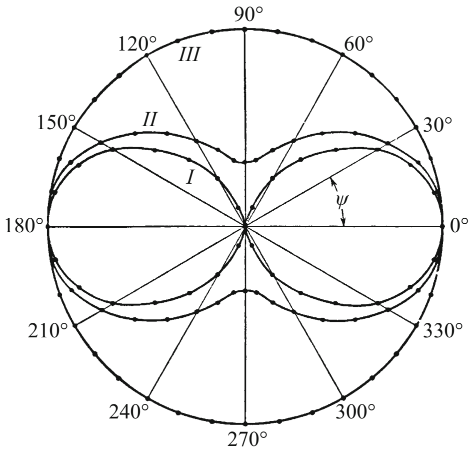


Figure 24.17 The radiant power (relative values) transmitted by the analyzer in Fig. 24.16, represented as the length of the radius. ψ is the angle between the plane of oscillation of the analyzer and that of the polarizer. Curve *I* refers to linearly-, *II* to elliptically-, and *III* to circularly-polarized light

of $E \cos \psi$. The transmitted radiant power must therefore be proportional to $\cos^2 \psi$. This agrees with the measurement; the results are shown graphically using polar coordinates in Fig. 24.17, curve *I*.

The zero values appear at $\psi = 90°$ and $= 270°$. This means that two “crossed” polarizers (*P* and *A*) allow no light to pass through from the lamp to the observer.

2. A mica platelet of thickness $d = 0.154$ mm is now inserted. It produces a path difference of $\varDelta = \lambda$, according to Eq. (24.2). The light remains linearly polarized, and we again measure curve *I*. The same holds for mica platelets whose thickness is an integral multiple of the above thickness, and thus giving path differences of $\varDelta = 2\lambda$, 3λ etc.

3. The mica platelet is 0.077mm thick, $\varDelta = \lambda/2$. We again obtain a curve of the form *I*, however rotated by 90°. At $\psi = 0°$ and $\psi = 180°$, no light is transmitted. The light is thus again linearly polarized, but its plane of oscillation relative to the polarizer *P* is tilted by 90° (not shown in Fig. 24.17).

4. The mica platelet is 0.038 mm thick, $\varDelta = \lambda/4$ (a “$\lambda/4$ plate” or “quarter-wave” plate). The deflection of the radiometer is independent of ψ, and we observe curve *III*. The light is circularly polarized.[C24.6]

5. The mica platelet has a thickness of $d = 0.167$ mm. $\varDelta = (1\frac{1}{12})\lambda$, equivalent to $d = \frac{1}{12}\lambda$. The light is elliptically polarized, we measure curve *II*, and the analyzer allows some light to pass for every angle ψ. At $\psi = 90°$ and $\psi = 270°$, there are more or less flat minima, but no longer is the transmitted intensity zero, as it would be with linearly-polarized light.

C24.6. There are also polarization foils available today for producing circularly-polarized light (Sect. 24.3); they convert linearly-polarized light into circularly-polarized light. They can also be obtained as combinations of linear and circular polarizing foils (keep in mind the direction of the light!).

6. Thus far, we have kept the amplitudes of the two sub-beams constant and have varied their path differences. Now, we keep the path difference constant at $\lambda/4$, i.e. we use a $\lambda/4$-plate and vary the ratio of the amplitudes. To do this, we change the angle between the plane of oscillation of the polarizer P and the vertical (i.e. the β direction of the mica platelet). In this way, we can produce elliptically-polarized light with any desired oscillation form using only a single mica platelet. We can thus obtain all of the curves as measured in Fig. 24.17 as well as their intermediate forms.

To conclude this section, we replace the red-filter light which we have used throughout the section by everyday incandescent light. Furthermore, we remove the radiometer and observe the images directly on a screen. The constant in Eq. (24.2) has a different value for every wavelength region; thus for example, with green light of wavelength $\lambda = 535$ nm (from a thallium-vapor lamp), we find

$$\frac{\varDelta}{\lambda} = 7.1\,\frac{d}{\text{mm}}\,. \tag{24.3}$$

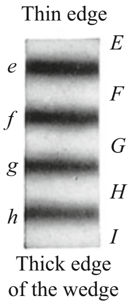
The individual wavelength ranges thus exhibit different path differences and polarization states. The analyzer allows some spectral ranges to pass, others less or not at all, i.e. for one range, curve *I* in Fig. 24.17 applies; for another curve *II* applies, etc. As a result, the image of the mica platelet appears in a variety of colors, which glow brightly for some thicknesses of the crystal.

24.6 The Interference of Parallel Beams of Polarized Light

In the last experiments discussed above, we superposed two coherent, transverse waves which were however oscillating in perpendicular planes, with arbitrary path differences. This yielded elliptically-polarized waves (including the limiting cases of linear and circular polarization), but no interferences, i.e. no changes in the spatial distribution of the waves, no maxima or minima as for example in Fig. 20.10. In order to produce "interference fringes", the *coherence* of the two beams alone is not sufficient; instead, they must also have a common plane of oscillation.

A common plane of oscillation can always be obtained by inserting an *analyzer* (e.g. *A* in Fig. 24.16). An analyzer permits only those components of the two waves oscillating perpendicular to each other that are parallel to its own plane of oscillation to pass through. In Fig. 24.16, the plane of oscillation of the polarizer and that of the analyzer are perpendicular to each other. They could also be set to be parallel; then all the maxima and minima in the interference patterns would exchange places. Take note of both possibilities. Following these general preliminary remarks, we will present two examples – in this section, only for parallel beams of polarized light.

Figure 24.18 Equidistant interference fringes in a quartz *wedge* cut parallel to the optical axis (collimated beam of red-filter light, length of the wedge 38.5 mm, thickness decreasing from 0.79 to 0.48 mm; photographic positive, just as in Fig. 24.19)



1. The mica platelet G in Fig. 24.16 is replaced by a long, flat *wedge* cut from a birefringent crystal (e.g. quartz). The direction denoted as optical axis is parallel to the edge of the wedge (Fig. 24.8, right), and this edge is placed horizontal. The radiometer M is now superfluous and can be removed. On the screen, using red-filter light we see the image of the wedge as photographed in Fig. 24.18. It exhibits interference fringes parallel to the edge of the wedge. Explanation: The interference fringes are curves of constant path difference. The crystal produces two sub-beams due to its birefringence. Their path difference depends on the thickness of the layer that they pass through. The interference fringes are thus a kind of curves of constant thickness. At the positions e, f, g etc., the path difference of the two sub-beams is equal to an integral multiple of the wavelength, so that $\varDelta = m \cdot \lambda$. Therefore, the light behind the birefringent crystal is polarized just as before entering it. It cannot pass through the analyzer; the fringes e, f, g etc. appear as deep black minima. The maxima E, F, G etc. occur for path differences of $\varDelta = (m \cdot \lambda + \lambda/2)$. The light is again linearly polarized behind the birefringent crystal, but its plane of oscillation is tilted by 90° and is now parallel to that of the analyzer. In the transition regions between e and E, f and F etc., the light is elliptically polarized. The analyzer allows some part of the light to pass through, depending on the form of the ellipse (compare Fig. 24.17).

Using ordinary incandescent light, the interference fringes appear as color-shaded bands. This is because the spacing of neighboring interference fringes is reduced as the wavelength decreases. Thus, with incandescent light, the interference fringes of the various wavelength regions overlap in different locations. This is true of all interference phenomena.

2. A *plane-parallel* quartz plate, about 1 mm thick, also cut parallel to its optical axis, is placed between the analyzer and the polarizer, whose planes of oscillation are set to be perpendicular to each other. The image of the quartz plate in incandescent light shows over its whole length the same chromatic hue as the wedge where it had the same thickness. We now cast the image of the plate not onto a screen, but rather onto the entrance slit of a spectrometer, and observe its spectrum on the screen. The spectrum is striped *across* its long di-

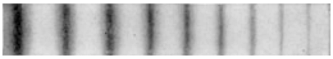


Figure 24.19 Interference fringes in a continuous spectrum, produced by a plane-parallel quartz *platelet* cut parallel to its optical axis and ca. 1.1 mm thick; here shown as a function of wavelength

rection by dark interference fringes (Fig. 24.19). The missing waves in these dark bands remained linearly polarized at the right of the apparatus, behind the birefringent plate, just as they were at the left, before entering the plate. Therefore, they could not pass through the analyzer.

24.7 Interference with Diverging Beams of Polarized Light

Interference with divergent polarized light can be dependably produced in the focal plane Z of a lens. The light source must have a large area. It is expedient to make the optical path on the image side of the lens telecentric (Fig. 24.20, top). Then one requires only small birefringent crystal plates. The light beams belonging to the image points 1 and 4 are dotted in the figure. They penetrate the crystal plate, just like the light beams of all the other image points, with parallel boundaries. Furthermore, all the light beams pass through the polarizer and the analyzer, in this case two polarization foils (Sect. 24.3). The plane of oscillation of these two foils are perpendic-

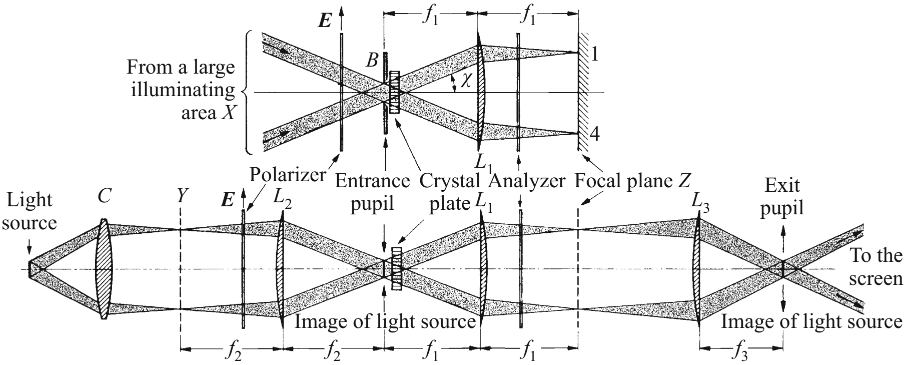


Figure 24.20 Top: Using *divergent* polarized light, interference fringes are produced in the focal plane of a lens. Bottom: The same as a demonstration experiment. The glowing surface X is an illuminated lens L_2. The image of the light source (a carbon arc) projected by L_2 acts as entrance pupil. In Z, there is not only the image of an infinitely distant plane, but also one of the plane Y determined by f_2. A *free-hand experiment*: Place the crystal platelet between two crossed polarization foils, hold it close to an arc lamp and observe the light on a screen.

ular to each other ("crossed polarizers"). The image plane Z is thus initially dark. Only after the birefringent crystal plate is inserted do we see in Z the image of a plane which is infinitely distant at the left. It shows interference fringes.

Examples:

1. A calcite plate, cut *perpendicular* to its optical axis (see Fig. 24.8, left), gives the interference pattern shown as a photograph in Fig. 24.21 a. It exhibits circular interference fringes and a dark cross. Explanation: The path difference of the two polarized sub-beams depends only on their angles of inclination χ (Fig. 24.20, top). Therefore, the curves of constant path difference, i.e. the interference fringes, are circular (they are thus a sort of "curves of constant inclination").

The crosses are interference-free regions. In them, there is only *one* polarized beam. The reason: We have drawn the crystal plate in Fig. 24.22 as seen from in front and enlarged. The numbers 1 and 4 mark the penetration points of the beam axes for the two light beams sketched in Fig. 24.20 (top). Furthermore the penetration points of three additional beam axes are shown. For each of them, the plane of incidence (a principal plane of the crystal) and the plane perpendicular to it are indicated by the dashed intersecting lines. The thick double arrows show the plane of oscillation of the light coming from the polarizer. Each beam is decomposed at the positions 2 and 3 into an ordinary and an extraordinary sub-beam. This is indicated by the thin double arrows. At the positions 1 and 4, in contrast, there is only an extraordinary beam, and at position 5, only an ordinary beam. One beam alone can never produce interference. Therefore, the incident light remains unchanged and thus cannot pass through the analyzer; the corresponding areas in the image remain dark.

2. A thick, uniaxial crystal plate, cut *parallel* to its optical axis, shows the interference pattern photographed in Fig. 24.21 b. It is visible only with monochromatic light (e.g. from a sodium-vapor lamp). With incandescent light, the orders of the interference fringes are too high. The curves of constant path difference have a hyperbolic shape. The detailed explanation of this phenomenon would take us too far afield here.

In Fig. 24.21 b, the path difference $\varDelta$ in the center of the image was equal to $m\lambda$. For $\varDelta = (m + \frac{1}{2})\lambda$, the dark and the light regions are exchanged. With a parallel beam of light in earlier figures (Fig. 24.16), we observed only the center part of this image, using the same plate.

3. A uniaxial crystal plate, cut at 45° to the optical axis, gives practically linear interference fringes. They could be considered to be extensions of the branches of the hyperbola in Fig. 24.21b.

4. We put two such plates together and rotate one of them by 90°. Then we obtain the complicated interference pattern photographed in Fig. 24.21 c. In incandescent light, one of the middle fringes appears colorless; it is thus due to rays with a zero path difference, and is a fringe of zeroth order. Its

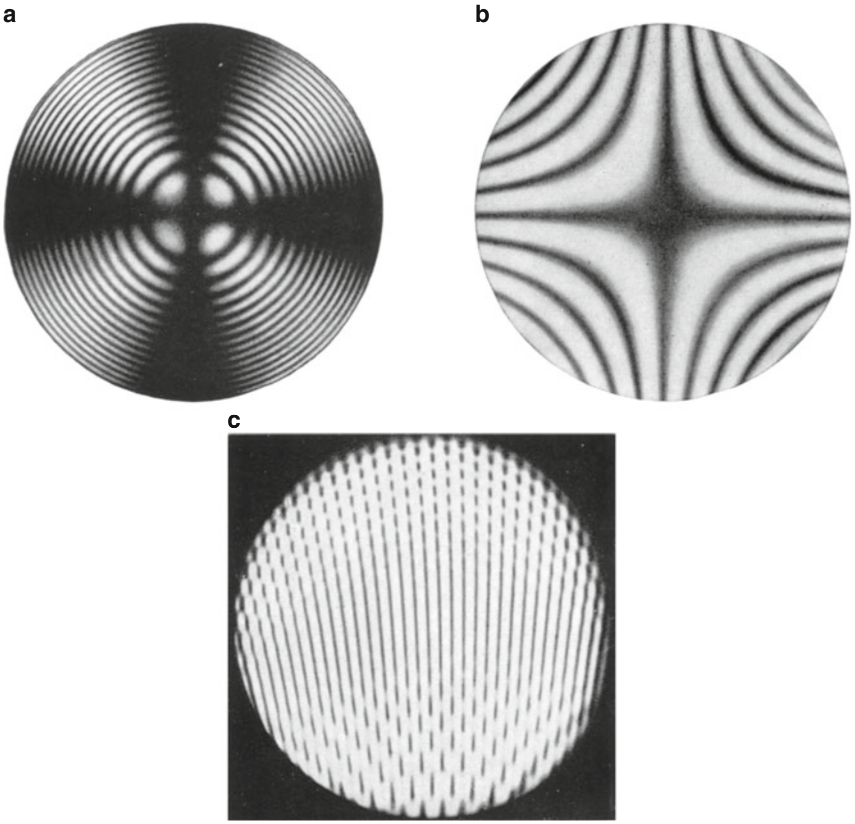


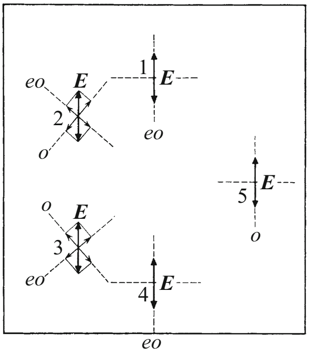
Figure 24.21 Three interference patterns of uniaxial crystals in divergent polarized light, photographed in the image plane Z of the apparatus in Fig. 24.20 (photographic positives). Part a shows a calcite plate ($d = 2\,\text{mm}$) cut perpendicular to its optical axis (using circularly-polarized light, the black cross can be eliminated). b: A quartz plate cut parallel to its optical axis ($d = 9\,\text{mm}$, Na-D light). c: Two quartz plates, cut about 45° to their optical axes and put one above the other, rotated by 90° (Savart's double plate).

All of the interference patterns obtainable with crystals and polarized light are noticeable for their strong intensity. This is a result of the coherence condition (Eq. (20.1)). Compare e.g. the interference pattern in Fig. 24.21 a with the one in Fig. 20.10. There, the angle 2ω was already very small; when using polarized light, it is zero. This means that both of the "rays" in each pair which can interfere with each other have the *same direction*. They nevertheless have a path difference, since they are polarized perpendicularly to each other and propagate in the material at different velocities. For $\sin 2\omega = 0$, one can use light sources of arbitrarily large diameter and thus obtain high light intensities.

neighboring fringes on both sides are colored, and the rest of the structure of the interference pattern is not visible with incandescent light.

F. Savart set two such crossed quartz plates, cut at 45° to their optical axes, together with a polarization prism in the same mount, and thus obtained a very sensitive *polarimeter*. It serves in many kinds of observations to detect small admixtures of polarized light in natural light. If you look at

Figure 24.22 The explanation of the dark cross in Fig. 24.21a



the sky or some illuminated object through such an instrument and rotate it around its long axis, then you will always see the interference fringes of low order, the colorless center fringe with its colored neighbors. A small fraction of the light is nearly always polarized; completely unpolarized light is an ideal limiting case.[C24.7]

C24.7. An historical note: Further details, in particular on the analysis of elliptically-polarized light, were given in the 13th edition of "*Optik und Atomphysik*", Chap. 10. Some relevant references in English can be found at https://www.osapublishing.org/josa/abstract.cfm?uri=josa-65-3-352 and at https://www.osapublishing.org/josa/abstract.cfm?uri=josa-50-9-892 .

24.8 Optically-Active Materials – Rotation of the Oscillation Plane. The FARADAY Effect

We now return to Fig. 24.16 and replace the mica platelet G by a quartz plate which is cut perpendicular to the optical axis. With this setup, a new phenomenon can be observed: The quartz plate *rotates* the plane of oscillation of the light. The angle of rotation α is proportional to the thickness d of the plate, that is

$$\alpha = \text{const} \cdot d\,. \tag{24.4}$$

The value of the constant for red-filter light is 18°/mm, and it increases strongly with decreasing wavelength. Therefore, if we use incandescent light instead of red-filter light, there is no position of the analyzer for which no light is transmitted; instead, at each position, we see a bright field of view which has a different color.

For demonstration experiments, a quartz plate 3.75 mm thick is particularly suitable. Two of these plates can be set up adjacent to each other, one of them made of right-rotating ("dextrorotatory") quartz, the other of left-rotating ("levorotatory") quartz[6]. Such a "sensitive double plate" exhibits a uniform purple color only when it is

[6] These are two mirror-image forms of quartz which rotate the light in either the one or the other direction ("right- and left-rotating", enantiomorphism; see for example Bergmann-Schaefer, "*Lehrbuch der Experimentalphysik*", Vol. 3, 10th edition (2004), Sect. 4.9). English: See for example http://www.quartzpage.de/gen_phys.html .

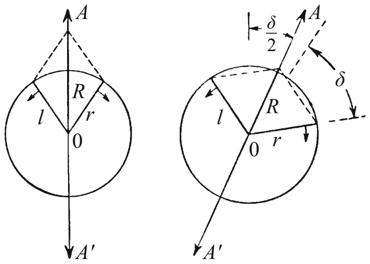


Figure 24.23 The superposition of two oppositely-rotating circular oscillations of the same frequency and amplitude. The direction *r* in the left-hand image is shown as the long dashed line outside the circle in the right-hand image.

placed between strictly parallel-oriented NICOL prisms. Even with very small angular deflections, the color hue on the one side changes towards the red, and on the other side towards the blue. This setup can be used to align the oscillation planes of polarizer and analyzer strictly parallel to each other, which is useful for example in calibrating measurement instruments, e.g. the saccharimeters which will be described below.

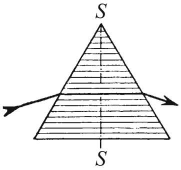
The ability to produce optical rotation (rotations of the oscillation plane of light), usually termed *optical activity*, is not specific to a crystalline structure of the material. It can also be found in molecules in solutions, for example for sugar dissolved in water. The angle of rotation of the plane of oscillation is in this case proportional not only to the layer thickness of the sample, but also to the concentration of the solution. Therefore, we can determine unknown concentrations from the value of the angle of rotation measured (for example in "saccharimeters"). Sugar molecules can also be right-rotating ("right-handed" or *dextrorotatory*) or left-rotating ("left-handed" or *levorotatory*). A 50 % mixture of the two is called a "racemic mixture".

Every linearly-polarized oscillation can be treated as the superposition of two circular oscillations of the same frequency and amplitude, but with opposite senses of rotation. In Fig. 24.23, at the left, l denotes the vector which is rotating to the left, and r the vector which is rotating to the right, while R is the resultant (sum) vector. Its end point passes along the double arrow AA'. The half-length OA is the amplitude of the linear oscillation (that is, the maximum value of its deflection). At the right, the same superposition is drawn, but now the oscillation of the vector rotating to the right leads that of the left-rotating vector by a phase difference δ. As a result, the resultant linear oscillation vector R is rotated by the angle $\delta/2$ in a clockwise direction.

Applied to the case of light, this means that a dextrorotatory material transmits a right-circularly-polarized light wave (see Fig. 24.15) more quickly than a left-circularly-polarized wave. The right-circular wave propagates more rapidly in the material than the other wave; it has a smaller refractive index. An optically-active material shows a new type of birefringence, called *circular birefringence*: It splits natural light not into two linearly-polarized beams (as in *linear birefringence*), but instead into two *circularly-polarized beams*.

This strange form of birefringence can be seen in all spectrometers which use simple quartz prisms. During the fabrication of such prisms, the line of symmetry SS (Fig. 24.24) is chosen to be perpendicular to the long

Figure 24.24 A quartz prism that exhibits birefringence in the dashed direction, which is denoted as its *optical axis*



direction of the quartz column, that is perpendicular to the optical axis. Nevertheless, we can see that all the spectral lines are split into two closely-spaced double lines. These two lines are circularly polarized in opposite directions.

The magnitude of this circular birefringence is very small. The refractive indices differ for example at $\lambda = 436$ nm by only 7 units in the fifth place after the decimal point. We can therefore in general simply define the optical axis to be the direction which is free of birefringence, in quartz and likewise in calcite, and in all the other non-optically-active birefringent crystals.

Because of the small magnitude of this circular birefringence, it is not suitable for demonstration experiments. For individual observations, the blue spectral line from a mercury-arc lamp is suitable. We place a $\lambda/4$-plate of mica in front of the ocular, together with a polarization analyzer. Then, depending on the positions of the β and the γ axes, we can see that one of the two spectral lines vanishes.

Paramagnetic and especially ferromagnetic materials rotate the plane of oscillation of light when they are placed in a magnetic field and the direction of light propagation is parallel to the field direction; this is the FARADAY *effect*, i.e. a rotation of the plane of the linearly-polarized light beam.[C24.8] Looking perpendicular to the field direction, we observe *birefringence*, with the optical axis parallel to the field direction.

[C24.8] See for example the book by Y.R. Shen, "Principles of nonlinear optics" (Wiley-Interscience, New York 1984) (http://www.osti.gov/scitech/biblio/6102640) for a detailed discussion of the FARADAY effect and other magneto-optical and electro-optical effects. See also https://en.wikipedia.org/wiki/Faraday_effect for a brief history and references.

24.9 Strain Birefringence. Conclusions

"Among the solid materials, there are numerous birefringent substances, including the crystals of all the non-regular systems, but a strictly single-refracting material can only be approximated".

In the field of *electromagnetism*, we distinguish between conductors and insulators. There are numerous conductors among the solid materials (in particular the metals), but a perfect insulator remains an idealized limiting case. In *optics*, we find a similar situation with the division into single- and double-refracting (birefringent) substances. Among the solid materials, there are numerous birefringent substances, including the crystals of all the non-regular systems, but a strictly single-refracting material can only be approximated. If we place thick layers (of several centimeter thickness) of apparently singly-refracting materials (regular crystals, glasses, transparent plastics) between crossed polarizers, e.g. in place of the plate G in Fig. 24.16, then we find that the field of view always shows brighter and darker regions, which are colored when incandescent light is used: The materials are birefringent in many more-or-less extended regions.

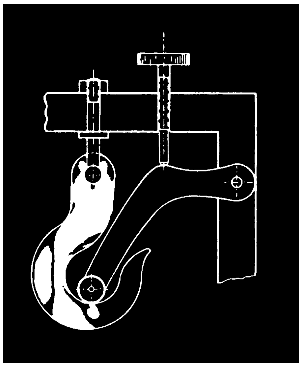


Figure 24.25 Strain birefringence in a model of a hook for a crane (the polarizers are perpendicular to each other and tilted by 45° to the vertical; photographic positive). (The holder, the lever to provide the load, and the outline of the hook were drawn into the photo to make them clearly visible)

This birefringence comes about through internal strains which vary from place to place within the material. Their practical elimination is tedious and expensive; the material must be heated to almost its melting point and then cooled very slowly. Glass blanks for large astronomical lenses have to be cooled over a period of many months. "Tempered" glass approaches the optical ideal of a solid object without birefringence rather closely. However, it must be carefully protected against mechanical stresses. Even pressing between the fingertips produces a noticeable birefringence.

For optical technology, strain birefringence is a source of annoying disturbances. For another technical field, however, namely the science of materials strength, it is extremely useful. With its help, the distribution of compressive and tensile stresses can be investigated in model experiments. For example, Fig. 24.25 shows a model profile of the hook for a crane, made of plastic and placed between two crossed polarizers. A load is applied via a lever. The regions which are stressed due to compressive or tensile strains appear brighter in the image. The dark boundary stripes between them show the strain-free transition regions, the "neutral fibers". The quantitative evaluation of such images is not simple. It is treated in the voluminous technical literature on this subject.

Our treatment of polarization has been limited to experiments using visible light. In the ultraviolet and infrared spectral ranges, one finds nothing new. Polarizers for the ultraviolet are described in Fig. 24.9, and for the infrared, they will be discussed later in Sect. 25.6. It is also more expedient to treat the polarization of X-rays later; it requires special experimental techniques (Sect. 26.8).

The Relation Between Absorption, Reflection and Refraction of Light

25

25.1 Preliminary Remark

In this entire chapter, we assume that we are dealing with *collimated beams* (i.e. parallel-bounded beams) of light, that is practically pure plane waves. This radiation is presumed to be monochromatic (single-frequency); for measurements, we use individual spectral lines, such as those from a metal-vapor lamp. In all the experiments presented, the plane of incidence (also known as the plane of reflection; see Sect. 16.3) of the light lies in the plane of the page. The amplitude of the light oscillating within that plane is denoted by $E_\parallel$, and that of light oscillating perpendicular to the plane by $E_\perp$.[C25.1]

C25.1. The quantity $\boldsymbol{E}$ denotes the vector of the electric field of the electromagnetic waves which constitute light, as in the previous chapter. $E_\parallel$ and $E_\perp$ are the components of the field parallel and perpendicular to the plane of incidence, and, as always, both are perpendicular to the propagation direction of the light.

25.2 The Extinction and Absorption Constants

In all of our observations thus far, we have assumed that the radiation was not attenuated on passing through a layer of matter. In that case, we needed only a single materials constant, namely the refractive index n of the material. If there is however an attenuation, then we need a second materials constant, the so-called *extinction constant* K (or an equivalent quantity derived from it). It is defined by a measurement procedure, just like the refractive index:

In Fig. 25.1, a collimated light beam is incident on a radiometer. Along its path, one of two sheets of the same material (but with different thicknesses, x_1 or x_2) are alternately placed in the beam. The difference of the thicknesses $\Delta x = (x_2 - x_1)$ is chosen to be small compared to the thickness x_1. The deflections S (the 'signal') of the radiometer give a relative measure of the radiant power $\dot{W}$ of the light arriving at the radiometer. These powers $\dot{W}_1$ and $\dot{W}_2$ (or the corresponding radiant *intensities*; see Eq. (19.2)) are in both cases smaller *with* the sheets than without them. This is for two reasons: First, some fraction of the radiation is lost to reflection at the front and rear surfaces of the sheets. These fractions are the same for both sheets.

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_25

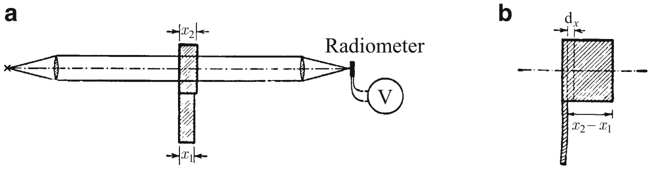


Figure 25.1 a: The definition of the *extinction constant* K; in the absence of scattering out of the beam, K is also called the *absorption constant*. b: Its measurement with thick absorber layers

Second, a fraction of the transmitted radiation is either “absorbed” (i.e. taken up by the material, that is converted into heat, chemical or electrical energy[5]), or it is “scattered” out of the beam. The fraction of the radiant power which *penetrates* into the sheet and is thus removed from the beam by *extinction* is larger for the thicker sheet than for the thinner one. The measurements yield

$$\left.\begin{aligned}(S_1 - S_2) &= \text{const} \cdot S_1 \Delta x\,,\\ \Delta\dot{W} &= \dot{W}_1 - \dot{W}_2 = K \cdot \dot{W}_1 \Delta x\,.\end{aligned}\right\} \tag{25.1}$$

In words, this means that the radiant power $\Delta\dot{W}$ that is lost by a collimated beam of light within a layer of material due to absorption and scattering is proportional to the power $\dot{W}_1$ which *penetrates* through the layer of material, and also to the thickness of the layer, $\Delta x = d$.[C25.3] The proportionality factor K is termed the *extinction constant*. If scattering plays no role in comparison to absorption, then we will call the extinction constant simply the *absorption constant*. If, in contrast, we can neglect absorption relative to scattering, then we will call it the *extinction constant from scattering*. The use of the concepts ‘extinction’, ‘extinction constant’, etc. by themselves leaves it open as to which relative importance is to be attributed to absorption and to scattering.

C25.2. Pohl gave a detailed treatment of fluorescence and phosphorescence in the 13th edition of “*Optik und Atomphysik*”, Chap. 15. For a modern discussion in English, see for example https://en.wikipedia.org/wiki/Phosphorescence .

C25.3. The proportionality between the absorbed radiant power and the layer thickness, and the exponential law which results (Eq. (25.2)), are often referred to in the literature as “Lambert’s law of absorption” (or “Lambert-Beer’s law”). (Lambert’s *cosine law* was treated previously in Sect. 19.2).

Equation (25.1) serves to *define* the extinction constant. For its practical *measurement*, the thickness difference $(x_2 - x_1)$ is nearly always chosen to be of the same magnitude as the layer thickness d; that is, not small compared to the thickness, as assumed above (Fig. 25.1b). We imagine the segment $(x_2 - x_1)$ to consist of thin layers dx and then sum over the absorption of all of these layers, obtaining

$$\int\limits_{\dot{W}_2}^{\dot{W}_1} \frac{\mathrm{d}\dot{W}}{\dot{W}} = \int\limits_{0}^{d} K \cdot \mathrm{d}x\,, \quad \text{thus} \quad \ln \dot{W}_1 - \ln \dot{W}_2 = K \cdot d\,, \quad \text{and}$$

$$\dot{W}_2 = \dot{W}_1\, e^{-Kd}\,. \tag{25.2}$$

The measurement of large extinction constants ($K > 10^4\,\text{mm}^{-1}$) is difficult. It requires extremely thin layers. In such layers, interference occurs,

[5] These energy forms can later be converted back into radiation (fluorescence – see Sect. 16.10 – and phosphorescence).[C25.2]

and furthermore, the reflectivity depends on the layer thickness. These difficulties can be avoided by using the following procedure: We first measure the ratio of incident to transmitted radiant power $(\dot{W}_\mathrm{i}/\dot{W}_\mathrm{t})$ as a function of the layer thickness d. Then we plot $\ln(\dot{W}_\mathrm{i}/\dot{W}_\mathrm{t})$ graphically against d. The larger values of $(\dot{W}_\mathrm{i}/\dot{W}_\mathrm{t})$ will lie on a straight line. Its slope is the extinction constant that we are seeking.

25.3 The Mean Penetration Depth w of the Radiation. The Extinction and Absorption Coefficient k

In this section, we first list some values of the absorption constants of various materials for light waves in the visible spectrum. They are set out in the third column of Table 25.1.

The reciprocal of the absorption constant K (or, more generally, the extinction constant), has an intuitively clear meaning: Along the path $w = 1/K$, the radiant power of a collimated light beam decreases to $1/e = 1/2.718 \approx 37\,\%$ of its initial value. This distance w will be called the *mean penetration depth* of the light. Examples of this useful quantity can be found in the fourth column of Table 25.1.

The *transparency* (everyday language) of a layer of material of thickness d depends on the ratio d/w. The smaller this ratio, the more transparent is the layer. Thus, at a thickness of a few µm, even pitch becomes transparent($w \approx 7\,\mu$m); and at a thickness around a hundred times smaller, metals are also transparent ($w \approx 10\,$nm).

For *wave phenomena, the wavelength is always the appropriate length scale, in particular the wavelength in vacuum* (or air). We thus use the ratio of the wavelength λ in vacuum (air) to the mean penetration depth w of the radiation in the particular material, i.e. λ/w.

C25.4. In carefully purified silicate glasses, for infrared light ($\lambda = 1.55\,\mu$m), absorption constants of $3.6 \cdot 10^{-8}\,\mathrm{mm}^{-1}$ can be attained, and thus mean penetration depths of around 28 km. With the aid of intermediate amplifiers, this makes it possible to manufacture and use optical-fiber cables for transmitting data around the globe (see Comment C16.10).

Table 25.1 Absorption constants, mean penetration depths, and absorption coefficients of some materials

Material	Wavelength λ in nm	Absorption constant K in mm^{-1}	Mean penetration depth of light $w = 1/K$	$\dfrac{\text{Penetration depth } w}{\text{Wavelength } \lambda}$	Absorption coefficient $k = \dfrac{1}{4\pi} \cdot \dfrac{\lambda}{w}$
Water	770	0.002_4	42 cm	550 000	$1.4 \cdot 10^{-7}$
Heavy flint glass[C25.4]	450	0.004_6	22 cm	500 000	$1.6 \cdot 10^{-7}$
“Black” neutral glass	546	10	0.1 mm	180	$4.4 \cdot 10^{-4}$
Pitch	546	140	7 µm	13	$6 \cdot 10^{-3}$
Brilliant green	436	7000	0.14 µm	0.32	0.25
Graphite	436	20 000	0.05 µm	0.11	0.72
Gold	546	80 000	0.01 µm	0.02_2	3.6

To simplify later trigonometric calculations, we multiply by $1/4\pi$ and define the resulting quantity as the *extinction coefficient* (or in the special case of no scattering, the *absorption coefficient*):C25.5

$$k = \frac{1}{4\pi}\frac{\lambda}{w} = \frac{1}{4\pi}K\lambda\,. \qquad (25.3)$$

C25.5. In some textbooks, the absorption constant K defined by Eq. (25.1) is also called the absorption *coefficient*.

Some values of k are shown in the last column of Table 25.1. Whether one uses the extinction quantity K or k in a particular case depends only upon which of the two makes possible a more convenient formulation of a statement.

25.4 BEER's law. The Interaction Cross-Section of a Single Molecule

Sometimes, one finds the extinction constant of a uniform material to be proportional to its density ϱ, or that of a solution to be proportional to its concentration c ("BEER's law"; see Fig. 25.2). In both cases, one can then define a *specific extinction constant*:

$$K_\varrho = \frac{K}{\varrho} \qquad (25.4)$$

and

$$K_\mathrm{c} = \frac{K}{c}\,; \qquad (25.5)$$

(here, ϱ is the density and c is the concentration, i.e. the amount of substance n of the dissolved molecules/volume V of the solution).

Example: From the slope of the line in Fig. 25.2, for an aqueous copper sulfate solution we obtain the specific extinction constant[1]

$$K_\mathrm{c} = \frac{K}{c} = \frac{1.71\,\mathrm{cm}^{-1}}{1\,\mathrm{mol/liter}} = 1710\,\frac{\mathrm{cm}^2}{\mathrm{mol}}\,.$$

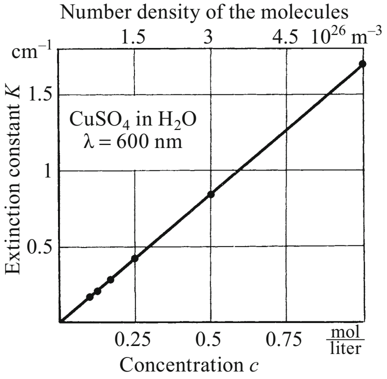
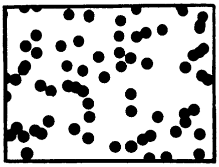


Figure 25.2 BEER's law and the measurement of the specific extinction constant K/c

[1] It is called the *molar* extinction constant in chemistry, as suggested by the units employed; this is however not completely consistent, since molar quantities are otherwise referred only to the amount of substance n.

Figure 25.3 A model experiment demonstrating the interaction cross-section of individual molecules



If the proportionality mentioned above is experimentally obeyed, i.e. when K/ϱ or K/c can be treated as constants, then the extinction takes place *without interactions between the individual molecules*. Then it is expedient in Eqns. (25.4) and (25.5), instead of the density ϱ or the concentration c, to use the

$$\begin{array}{l}\text{Number density}\\ \text{of the molecules } N_\text{V}\end{array} = \frac{\text{Number } N \text{ of active molecules}}{\text{Volume } V \text{ of the body or the solution}}$$

(as in Fig. 25.2, upper abscissa scale), that is:

$$K = K_\varrho \cdot \varrho = \frac{K_\varrho M_\text{n}}{N_\text{A}} \cdot N_\text{V} \quad \text{and} \quad K = K_\text{c} \cdot c = \frac{K_\text{c}}{N_\text{A}} \cdot N_\text{V}$$

(M_n is the molar mass $= M/n$, n is the amount of substance, and N_A is the AVOGADRO constant $= 6.022 \cdot 10^{23}\ \text{mol}^{-1}$).

The extinction constant K is the reciprocal of a length. Therefore,

$$\frac{K}{N_\text{V}} = \frac{KV}{N} \tag{25.6}$$

has the dimensions of an *area*. We call it the *interaction cross-section* σ of a molecule. In the limiting cases discussed in Sect. 25.2, σ is an "absorbing" or a "scattering" cross-section (i.e. the area of the "target" for absorption or scattering).

Example: From Fig. 25.2, we find the following value for the interaction cross-section:

$$\sigma = 2.82 \cdot 10^{-25}\ \text{m}^2\,.$$

The physical significance of the interaction cross-section can be explained in an intuitively apparent manner. Figure 25.3 shows a "snapshot" of a model gas composed of steel balls, with a layer thickness of 1 cm. We see the projection of the cross-sectional areas of the individual molecules. When brought into a collimated beam of radiation, each of these areas acts as if it were completely opaque; the radiation can continue in its original direction only in the gaps between the molecules. If we stack such layers of a model gas with a statistically-disordered distribution of molecules, then the overall area of the remaining free gaps decreases according to an exponential function, and this yields Eq. (25.2).[C25.6]

C25.6. Here, we thus have $\sigma = \pi r^2$. In general, we find

$$N_\text{V}\,\sigma\,l = 1\,,$$

where l is the 'mean free path'. For radiation, $l = w$, the mean penetration depth; for gases, it is the mean free path between collisions (compare Fig. 2.15).

Video 25.1:
"Absorption"
http://tiny.cc/sfggoy
The absorption process is demonstrated using a torsional-wave machine.

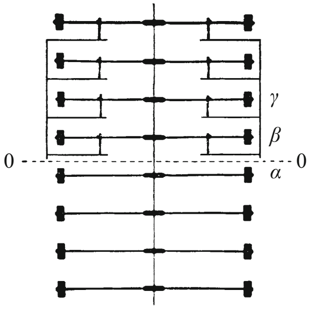


Figure 25.4 Some of the arms of a torsional-wave machine. The upper portion is fitted with an adjustable frictional damping mechanism (**Video 25.1**)

25.5 Distinguishing Between Weakly- and Strongly-Absorbing Materials

Distinguishing between weakly- and strongly-absorbing materials is of great importance for what follows in this chapter. This distinction is made by employing the mean penetration depth w of the radiation, or else the absorption constant K:

$$\begin{array}{cc} \text{Weak absorption means:} & \text{Strong absorption means:} \\ w = \frac{1}{K} > \lambda \quad \text{or}^{2} \quad k < 0.1 & w = \frac{1}{K} < \lambda \quad \text{or}^{2} \quad k > 0.1\,. \end{array} \tag{25.7}$$

"Seldom have physical terms been chosen in such a misleading way as the words 'weakly' and 'strongly' absorbing."

Seldom have physical terms been chosen in such a misleading way as the words "weakly" and "strongly" absorbing.

"Weakly" absorbing materials, for example diluted inks, can nevertheless absorb the entire incident radiant power or intensity (at a sufficiently large layer thickness d), apart from the minor losses due to reflection. "Strongly" absorbing materials, for example metals, in contrast, can absorb only a *small fraction* of the incident radiant power. The greater part cannot penetrate the material at all, and is reflected by its surface.

This is quite general, as can be demonstrated using mechanical waves. In Fig. 25.4, a short portion of a *torsional-wave machine* is sketched. This machine is fitted with a damping mechanism (above the dashed line $O - O$ in the figure); it consists of small brushes at the ends of the oscillating arms. The brushes scrape over rough paper surfaces when the machine is set into motion. These paper surfaces can be raised or lowered together, thus varying the frictional damping of the waves. Along the axis of this machine, we excite a short wave group ($\lambda \approx 60\,\text{cm}$), starting from below and travelling upwards. We can make three observations:

1. Without damping, the wave group passes over the boundary $O - O$ with no effect.

[2] If we round 0.08 up to 0.1.

2. With strong damping: The dumbbell arms β are strongly impeded by the damping. They can take up only a small fraction of the oscillation energy from the arms α. The major part is reflected, so that the amplitude of the wave group which travels back down is barely smaller than that of the original group which moved upwards.

3. The energy which is transferred to β in spite of the damping is mostly converted into frictional heat. A small remainder is passed on to γ, etc. Thus, the wave motion dies out over a short distance within the "absorbing material". Its mean penetration depth w in our example is only a small fraction of the wavelength λ. In the case of "strong" absorption, i.e. $w < \lambda$, *the waves cannot penetrate into the material. Only a small amount of energy is absorbed, and this occurs over a short distance* (cf. Sect. 25.8).

25.6 Specular Light Reflection by Planar Surfaces

Following our detailed treatment of the second optical materials constant, the *extinction constant* K or the *extinction coefficient* k, we will now discuss on an experimental basis the specular reflection of light from the planar surfaces of homogeneous materials.

In Fig. 25.5, a collimated light beam passes through a polarizer P, and the now linearly-polarized beam is incident on a radiometer, either directly (lower radiometer deflection $S_1 \propto \dot{W}_1$), or else after reflection by the surface M (upper radiometer deflection $S_2 \propto \dot{W}_2$). The plane of oscillation of the light is alternately chosen to be parallel ($E_\parallel$) or perpendicular ($E_\perp$) to the plane of incidence; furthermore, the angle of incidence α is varied (the limiting case of $\alpha = 0$, i.e. perpendicular incidence, can be only approximately achieved with this simple setup). The analyzer A is initially not present in the optical path. Each time, we measure the

$$\text{Reflectivity } R = \frac{\text{Reflected radiant power}}{\text{Incident radiant power}} = \frac{\dot{W}_2}{\dot{W}_1}\,. \qquad (25.8)$$

The amplitude of a light wave is proportional to the square root of its radiant power (or of the deflection of the radiometer).C25.7 Then, for the ratio of the amplitude E_r of the reflected light vector to the amplitude E_i of the incident light vector, we can write:

$$\frac{E_\mathrm{r}}{E_\mathrm{i}} = \sqrt{\frac{\dot{W}_2}{\dot{W}_1}}\,. \qquad (25.9)$$

The results of several measurements can be seen in Fig. 25.6 a–c. In Fig. 25.6 a and b, the mirror is made of materials with "weak" absorption (typically dielectric materials), while in Fig. 25.6 c, it is made of a material which exhibits "strong" absorption (typically

C25.7. Light waves are electromagnetic waves (see Comment C24.2). The statement that their amplitudes are proportional to the square roots of their radiant powers was dealt with in Comment C12.5.

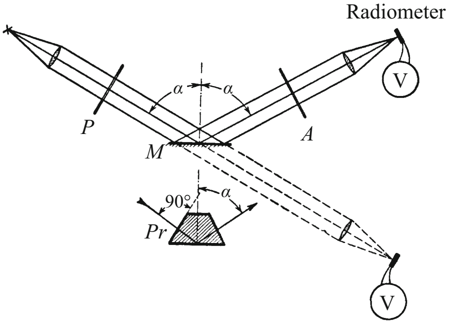


Figure 25.5 Measuring the reflectivity at various angles of incidence α (P: polarizer, A: analyzer). The plane of incidence is the plane of the page

a metal). Placing the results for these various cases side by side as in Fig. 25.6 a–c is the best way to illustrate their common characteristics and their differences. We want to point out in particular four aspects:

1. The ratio $E_\mathrm{r}/E_\mathrm{i}$ is much larger for strongly-absorbing materials than for weak absorbers *in the range of small and medium angles of incidence* α.

2. If the light vector lies parallel to the plane of incidence, then for weak absorption, there is a characteristic angle α_P. It is called the *polarization angle* or Brewster's angle, and it occurs for the following reason: When the incident light is unpolarized, at the angle of incidence α_P, only that fraction is reflected whose vector is perpendicular to the plane of incidence. Thus, the reflected light has become linearly polarized.

The French researcher E.L. MALUS was thus able to discover the linear polarization of light in 1808 by studying its reflection. Unfortunately, with this method we lose 84 % of the incident radiant power (Fig. 25.6 a). Furthermore, the kink in the optical path is inconvenient.
In the infrared, this method of polarization is indispensable even today. For example, at wavelengths greater than about 3 µm, one can utilize substances with very high refractive indices, e.g. selenium or lead sulfide, and thus avoid the large losses which occur in the visible spectral region. Mirror surfaces made from such materials can be prepared in a similar manner to most metallic surfaces: The material is evaporated in high vacuum and condensed onto a polished (and if necessary, cooled) glass plate.

3. At the polarization angle α_P, the reflected beam is perpendicular to the refracted beam. This is consistent with BREWSTER'*s law*:

$$\tan\alpha_\mathrm{P} = n \qquad (25.10)$$

(Derivation: $\sin\alpha_\mathrm{P} = n\,\sin\beta = n\,\sin(90^\circ - \alpha_\mathrm{P}) = n\,\cos\alpha_\mathrm{P}$
(β is defined as in Fig. 16.4)).

With Eq. (25.10), α_P can be utilized to measure the refractive index n.

4. With strong absorption, there is no polarization angle or BREWSTER angle α_P. Instead, the *principal angle of incidence* Φ is ob-

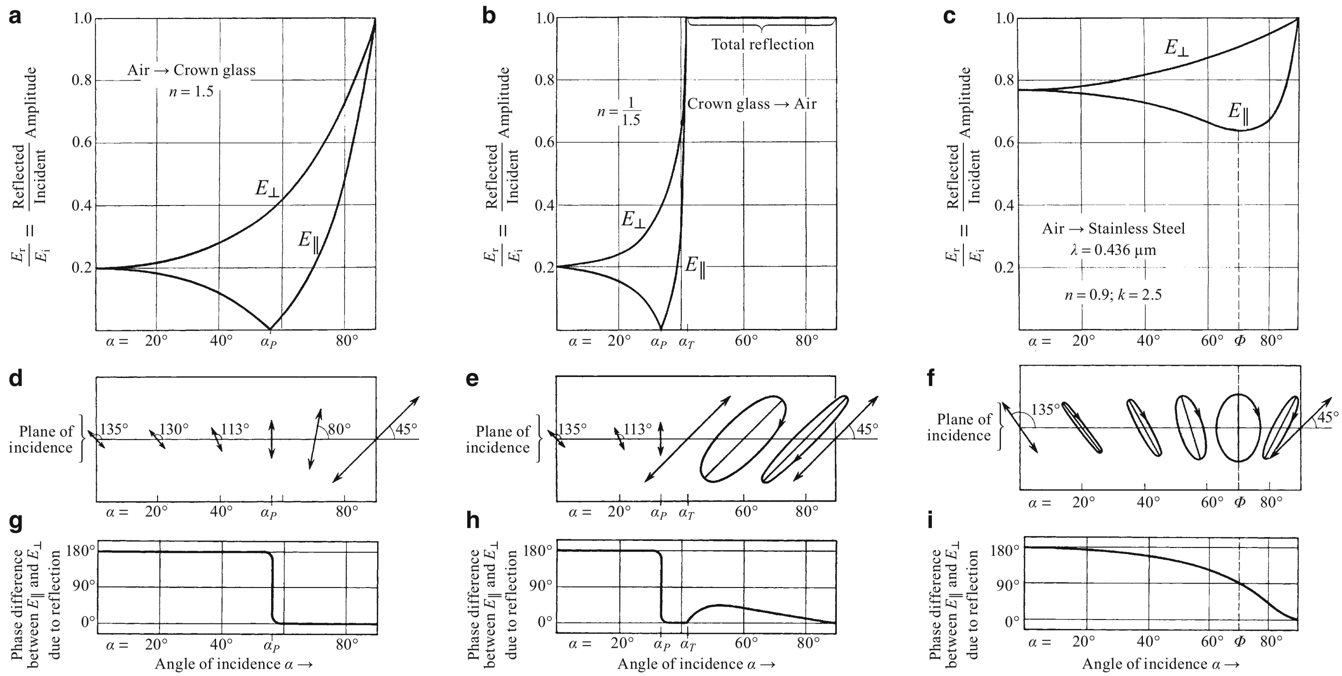


Figure 25.6 The influence of the angle of incidence on the reflection of linearly-polarized light. In the upper row of graphs, the light is reflected in graph a at the boundary between air and crown glass, and in graph b at the boundary between crown glass and air (to carry out these measurements, we require a prism *Pr* as sketched below in Fig. 25.5). In graph c, the light is reflected from the surface of a metal. The graphs in the center row, d–f, show the paths that would be traced out by the tip of the light vector for an observer looking at all angles of incidence *in the direction of light propagation* (not only for the incident beam, but also along the reflected beam). For these observations, the plane of oscillation of the incident light at $\alpha \approx 0°$ was inclined by 45° relative to the plane of incidence, as shown in Fig. 25.7. The graphs in the bottom row, g–i, show the phase differences which are required to describe the experimentally observed orbits in graphs d–f. These phase differences agree with those calculated from FRESNEL's formulas (and Eqns. (25.22) and (25.40), which are derived from them).

served (Fig. 25.6c). It can be employed when the two optical constants n and k are to be measured in strongly-absorbing materials (Sect. 25.12).

25.7 Phase Changes on Reflection of Light

We now no longer alternate between incident light beams which are oscillating perpendicular and parallel to the plane of incidence; instead, we fix the angle (called the azimuthal angle) between the light vector and the plane of incidence at a value of of $\psi = 45°$. This is sketched in Fig. 25.7 for $\alpha \approx 0$. The figure depicts a special case, drawn in perspective, of our general arrangement. The latter is shown in Fig. 25.8 without a perspective view, again making the plane of the page coincide with the plane of incidence. Our agreed-upon arrangement states: In every case, the positive directions of $E_{\|}$, $E_{\perp}$, and z follow each other in a similar sequence as the x, y, and z axes of a right-hand coordinate system. (In such a coordinate system, if we look along the z direction, we must rotate the x axis in a clockwise sense to bring it into the original direction of the y axis).

The experimental setup in Fig. 25.5 is now complemented by adding the polarization analyzer A. It can be rotated around the beam axis of the reflected beam. We then obtain measurement results as shown in Fig. 24.17. They are plotted in Fig. 25.6 d–i : Reflection produces not only different amplitudes $E_{\|}$ and $E_{\perp}$, but also *phase differences* between the vectors $\boldsymbol{E}_{\|}$ and $\boldsymbol{E}_{\perp}$. When these are not equal to 0° or 180°, the reflected light is elliptically polarized. With weak absorption, this

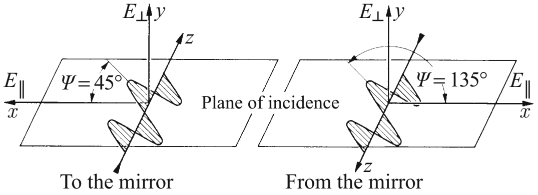
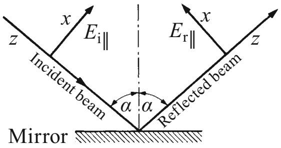


Figure 25.7 The orientation of the light vectors in the special case of nearly perpendicular light reflection and an observer who is always looking along the direction of light propagation

Figure 25.8 The orientation of the light vectors for an arbitrary angle of incidence α



$E_{i,\perp}$ and $E_{r,\perp}$ point upwards, perpendicular to the plane of the page occurs only in the region of total reflection (indicated in Fig. 25.6 b). For strongly-absorbing materials, however, it occurs at all angles of incidence.

At the *principal angle of incidence* Φ, the phase difference between $\boldsymbol{E}_{\parallel}$ and $\boldsymbol{E}_{\perp}$ becomes 90°. After two reflections at the principal angle of incidence Φ, the light is thus again linearly polarized. This is the basis for a convenient measurement method for Φ, useful also for demonstration experiments (J. JAMIN, 1849).

25.8 FRESNEL's Formulas for Weakly-Absorbing Materials. Applications

The entire empirical content of the left-hand and the center columns in Fig. 25.6 (graphs a and b, d and e, and g and h) was summarized by A. FRESNEL (1788–1827) in simple formulas. If we write the law of refraction as $\sin\alpha/\sin\beta = n$, then for the reflected radiation, we have:[C25.8]

$$\frac{E_{\mathrm{r}\perp}}{E_{\mathrm{i}\perp}} = -\frac{\sin(\alpha-\beta)}{\sin(\alpha+\beta)} \tag{25.11}$$

and

$$\frac{E_{\mathrm{r}\parallel}}{E_{\mathrm{i}\parallel}} = \frac{n\cos\alpha - \cos\beta}{n\cos\alpha + \cos\beta} = \frac{\tan(\alpha-\beta)}{\tan(\alpha+\beta)}\,. \tag{25.12}$$

For the radiation that penetrates into the material, we find

$$\frac{E_{\mathrm{p}\perp}}{E_{\mathrm{i}\perp}} = \frac{2\sin\beta\cos\alpha}{\sin(\alpha+\beta)} \tag{25.13}$$

and

$$\frac{E_{\mathrm{p}\parallel}}{E_{\mathrm{i}\parallel}} = \frac{2\sin\beta\cos\alpha}{\sin(\alpha+\beta)\cos(\alpha-\beta)}\,. \tag{25.14}$$

In the special case of perpendicular incidence, it follows from Eq. (25.11) for $\alpha \to 0$ that:

$$\frac{E_{\mathrm{r}}}{E_{\mathrm{i}}} = -\frac{n-1}{n+1}\,. \tag{25.15}$$

By squaring this equation, we obtain for *one* boundary surface the

$$\text{Reflectivity } R = \frac{\text{Reflected radiant power}}{\text{Incident radiant power}} = \left(\frac{n-1}{n+1}\right)^2 \tag{25.16}$$

[C25.8] POHL gave a derivation of FRESNEL's formulas in the 13th edition of "*Optik und Atomphysik*", Chap. 11. For a derivation in English, see for example http://physics.gmu.edu/~ellswort/p263/feqn.pdf which we introduced in Sect. 25.6; this is an important and often-used relation.

Examples: For glass, with $n = 1.5$, $R = 4\,\%$; for germanium, with $n = 4$, we find $R = 36\,\%$. The penetration of radiation can thus by no means be prevented by strong absorption alone.

From Eq. (25.16), it seemed for a long time that it would not be possible to manufacture reflection-free glass surfaces; however, making use of interference in thin evaporated crystal layers, a considerable degree of "anti-reflection" or "dressing" can be achieved (Sect. 20.12). In the first practically successful method, thin crystalline layers (e.g. of KBr or CaF_2) were evaporated onto quartz glass in high vacuum (G. BAUER, 1934).[C25.9]

C25.9. GERHARD BAUER, Dr. rer. nat. Göttingen 1931; *Annalen der Physik* **39**, 434 (1934).

The minus sign in Eq. (25.15) means that $\boldsymbol{E}_\mathrm{r}$ and $\boldsymbol{E}_\mathrm{i}$ are directed oppositely to one another for $n > 1$, and they are parallel for $n < 1$. The reflection produces a phase jump of 180° or $\lambda/2$[C20.3] for $n > 1$. When $n < 1$, in contrast, the phase remains unchanged.

The demonstration experiment of THOMAS YOUNG (1802): In a demonstration of NEWTON's rings (Sect. 20.9), he bounded the air layer by a piece of cambered glass with a small refractive index, and a piece of planar polished glass with a large refractive index. Then he filled a section of the air gap with a liquid whose refractive index lay between those of the two pieces of glass. In this region, the bright and dark interference fringes exchanged positions[3].

With our knowledge of this phase jump, we illustrate the perpendicular reflection at the planar surface of a weakly-absorbing material graphically for two examples in Fig. 25.9. For the perpendicular reflection, we use a single coordinate system whose z direction coincides with the direction of incidence of the light.

FRESNEL's formulas (25.13) and (25.14) hold for the light that penetrates beyond the boundary surface and into the material. It is expedient to illustrate them graphically (Fig. 25.10).

The amplitude ratio $E_\parallel/E_\perp$ for an oblique passage through the boundary does not reach its maximum value at the polarization angle $\alpha_\mathrm{P} = 56°19'$, but rather it continues to increase with increasing angle of incidence, as will be shown in the following.

When a collimated light beam passes at an oblique angle through a glass plate, one obtains partially-polarized light, i.e. a mixture of natural light and linearly-polarized light. Quantitatively, this light is characterized by its

$$\text{Degree of polarization } Q = \left|\frac{\dot{W}_{E\parallel} - \dot{W}_{E\perp}}{\dot{W}_{E\parallel} + \dot{W}_{E\perp}}\right| \tag{25.17}$$

($\dot{W}$ is the radiant power).

If we produce the partially-polarized light from a collimated light beam which passes at an oblique angle through a glass plate, then the degree of

[3] R. W. Pohl, *Physikalische Blätter* **17**, 208 (1961).

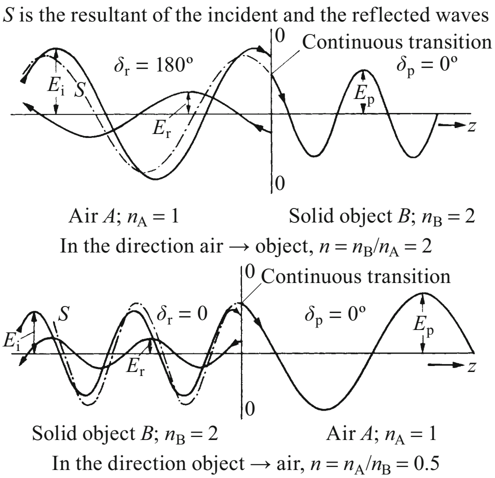

S is the resultant of the incident and the reflected waves
0
Continuous transition
$\delta_r = 180°$
$\delta_p = 0°$
E_i
S
E_p
E_r
z
0
Air A; $n_A = 1$
Solid object B; $n_B = 2$
In the direction air → object, $n = n_B/n_A = 2$
0
Continuous transition
S
$\delta_r = 0$
$\delta_p = 0°$
E_i
E_p
E_r
z
0
Solid object B; $n_B = 2$
Air A; $n_A = 1$
In the direction object → air, $n = n_A/n_B = 0.5$

Figure 25.9 Two examples of a perpendicular passage of travelling waves through the boundary $O - O$ between two materials of different refractive indices. 'Snapshot' images of this type continually change their form over time, but they repeat themselves periodically. In each snapshot, at each moment in time, the sum of the incident and reflected light vectors at the boundary is thus equal to the light vector of the light that penetrates the boundary.

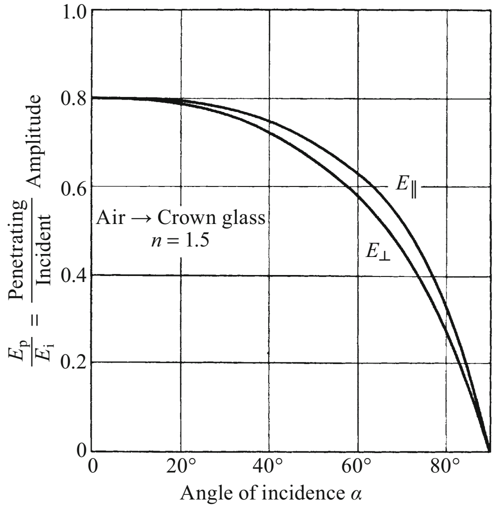


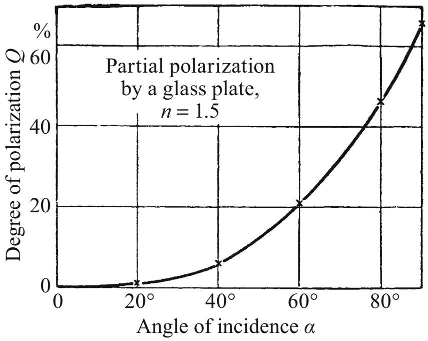
Figure 25.10 The penetration of light with the polarizations $E_\perp$ and $E_\parallel$ into an optically-denser material which absorbs weakly, corresponding to Eqns. (25.13) and (25.14)

polarization will be

$$Q = \frac{1 - \cos^4(\alpha - \beta)}{1 + \cos^4(\alpha - \beta)} \qquad (25.18)$$

(α is the angle of incidence, and $\sin\beta = \frac{1}{n}\sin\alpha$).

Figure 25.11 Influence of the angle of incidence on the degree of polarization of the light transmitted by a glass plate



The degree of polarization for a given refractive index n is thus determined by the angle of incidence α. Figure 25.11 shows an example for $n = 1.5$ which is practically important. From the continuous increase of the degree of polarization Q with increasing angle of incidence α, it follows that the amplitude ratio $E_\parallel / E_\perp$ also increases with α.
Derivation of Eq. (25.18): From Eqns. (25.13) and (25.14), we find for the passage through *one* boundary surface:

$$\frac{E_{p\parallel}}{E_{p\perp}} = \frac{1}{\cos(\alpha - \beta)} = a \quad \text{and through } two \text{ boundary surfaces} \quad \frac{E_{p\parallel}}{E_{p\perp}} = a^2 \,. \tag{25.19}$$

The radiant powers $\dot{W}$ are proportional to the squares of the amplitudes, that is

$$\frac{\dot{W}_{E\parallel}}{\dot{W}_{E\perp}} = a^4 \,, \tag{25.20}$$

and, from Eq. (25.17)

$$Q = \left| \frac{\dot{W}_{E\parallel} - \dot{W}_{E\perp}}{\dot{W}_{E\parallel} + \dot{W}_{E\perp}} \right| = \frac{a^4 - 1}{a^4 + 1} \,. \tag{25.21}$$

Inserting $a = 1/\cos(\alpha - \beta)$ yields Eq. (25.18).

25.9 Detailed Description of Total Reflection

In Fig. 25.12, a thin layer A with a refractive index n_A is sandwiched between two sheets of material B with planar surfaces and a larger refractive index, n_B. Waves are incident from the lower left at the angle of incidence α. They undergo total reflection when α surpasses the critical angle α_T for total reflection, as defined by Eq. (16.7), i.e. $\sin \alpha_\mathrm{T} = n_\mathrm{A}/n_\mathrm{B}$.

Total reflection can occur only when the thickness d of the layer of material A is at least of the order of the wavelength (Vol. 1, Sect. 12.9). Thinner layers are not an insurmountable obstacle for the waves; they can pass through the layer, although attenuated, as though they were passing through a tunnel: This is called the *tunnel effect*[C25.10], or "frustrated total internal reflection".

C25.10. Total reflection and the tunnel effect are discussed in detail in Vol. 1 using water waves as an example system (Sect. 12.9). In earlier editions, the series of images in Fig. 12.23, Vol. 1, could be found here in this chapter on optics.

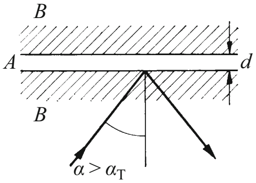


Figure 25.12 Avoidance of total reflection. The tunnel effect (**Video 12.2** from Vol. 1)

Video 12.2 from Vol. 1: **"Water-wave experiments"** http://tiny.cc/tfgvjy. The experiments on total reflection and the tunnel effect are shown after the time mark at 5:30 min.

For light, the demonstration is carried out using waves from the infrared spectral region. In Fig. 25.13, the crater of an arc lamp is imaged onto a radiometer M by two lenses made of rock salt. The collimated beam between the lenses is split into two beams by a mask B_1. An aperture B_2 which can be shifted in the vertical direction allows one or the other of the two sub-beams to pass. The two sub-beams are then incident on three 90° prisms made of rock salt. The bases of the small prisms are separated from the base of the large prism by small strips of metal foil, above with a thickness of 15 μm, below with 5 μm.

The visible part of the two sub-beams undergoes total reflection; it exits to the sides in the direction of the arrows. Likewise, the infrared radiation of the upper sub-beam undergoes total reflection. For the lower beam, in contrast, the radiometer indicates a large deflection. Radiation thus passes through the lower pair of prisms. This means that a 5 μm thick layer of air behind the base of the large prism avoids total reflection. But a 15 μm thick air layer allows the total reflection to occur without hindrance. As a result, the infrared radiation in the two beams contains waves of up to about 15 μm wavelength. (Waves of wavelengths longer than 15 μm were already absorbed by the first rock-salt lens. Details are given in Sect. 27.2).

This experiment using two prisms is also technically important. We can make the spacing of their basal areas variable; we then would have the possibility of changing the transmitted radiant power through tiny shifts in the spacing, i.e. of *continuously varying the intensity*. Furthermore, the two prisms can be used in the infrared spectral range as a filter. They stop shorter wavelengths and allow the longer ones to pass through.

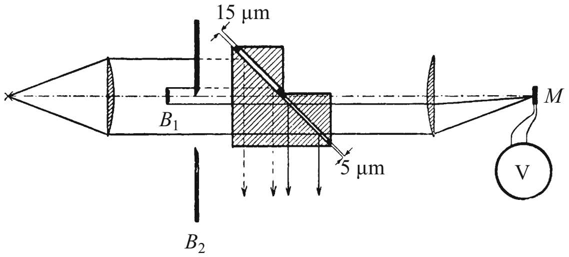


Figure 25.13 Demonstration of total reflection of infrared light and its avoidance by the "tunnel effect"

According to Fig. 25.6 h, in the region of total reflection, there is a phase difference δ between $\boldsymbol{E}_\parallel$ and $\boldsymbol{E}_\perp$. Thus, linearly-polarized light which has a component within the plane of incidence as well as perpendicular to it will be converted to elliptically-polarized light by reflection. We find (for $n < 1, \alpha > \alpha_\mathrm{T}$):

$$\tan\frac{\delta}{2} = \frac{\cos\alpha\sqrt{\sin^2\alpha - n^2}}{\sin^2\alpha}\,. \qquad (25.22)$$

Example: For $n = 1/1.5$, $\delta = 45°$ for two angles of incidence, at $\alpha = 48.5°$ and also at $\alpha = 54.5°$.
Derivation: The law of refraction $\sin\beta = \dfrac{1}{n}\sin\alpha$ yields values of $\sin\beta > 1$ only for $n < 1$. Then we find

$$\cos\beta = \sqrt{1-\sin^2\beta} = i\cdot\frac{1}{n}\sqrt{\sin^2\alpha - n^2}\,, \qquad (25.23)$$

an *imaginary* quantity ($i = \sqrt{-1}$). This is inserted into the FRESNEL formulas, and then the computation is carried out according to the same scheme as in Sect. 25.11.[C25.11]

C25.11. A detailed derivation of Eq. (25.22) can be found in Max Born and Emil Wolf, "Principles of Optics" (4th ed.), Pergamon Press (1970), Sect. 13 (available online at https://archive.org/details/PrinciplesOfOptics) .

25.10 The Mathematical Representation of Damped Travelling Waves

Travelling waves were treated in Sect. 12.1 of Vol. 1. Their phase velocity was denoted there by c. In optics, the phase velocity is the velocity of light c. Within a material of refractive index n, the phase velocity is reduced to c/n. In optics, an undamped travelling wave can be represented by the equation

$$E_\mathrm{x} = A\sin\omega\left(t - \frac{z}{c/n}\right) \qquad (25.24)$$

(E_x is the momentary value of the x component of the light vector (the vector of the electric field) at the location z and at time t; the wave is travelling in the positive z direction, and A is its amplitude, $\omega = 2\pi\nu$ its circular frequency, c/n = its phase velocity in a material, and n is the refractive index of that material).

We can carry out computations more easily with exponential functions than with trigonometric functions. Therefore, we replace the trigonometric functions by an exponential function, making use of EULER's relation:

$$e^{i\varphi} = \cos\varphi + i\sin\varphi\,, \quad i = \sqrt{-1}\,. \qquad (25.25)$$

Instead of Eq. (25.24), we can write

$$E_\mathrm{x} = A\,e^{i\omega(t - zn/c)}\,, \qquad (25.26)$$

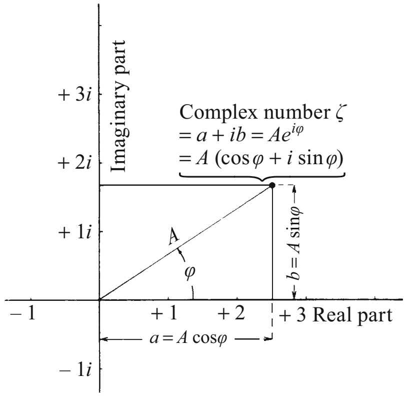


Figure 25.14 Graphical display of a complex number

and then we can use complex numbers for calculations and employ separately either the imaginary or the real part.

Complex numbers are pairs of numbers with particular *rules of calculation*, developed especially for such pairs. The words “imaginary” and “complex” are of only historical importance.
For the following sections, we need to note only a few items:
A complex number

$$\zeta = Ae^{i\varphi} = A(\cos\varphi + i\sin\varphi) = a + ib \tag{25.27}$$

(A is the “magnitude” and φ the phase angle of the complex number) can be represented graphically (Fig. 25.14):
To calculate the angle φ, we make use of the equation

$$\tan\varphi = \frac{\sin\varphi}{\cos\varphi} = \left.\frac{\text{Imaginary part}}{\text{Real part}}\right\} \text{ of the complex number } \zeta. \tag{25.28}$$

The “magnitude” A of a complex number $(a \pm ib)$ is found by multiplying it by its “complex conjugate” $(a \mp ib)$, so that for example

$$A^2 = (a + ib)(a - ib) = a^2 + b^2\,. \tag{25.29}$$

In these two equations, we find pure real numbers as the results. In other cases, one finds complex numbers on both sides of the equals sign, e.g.

$$a + ib = C + iB\,. \tag{25.30}$$

This then means that $a = C$ and also $b = B$ is a physical result, i.e. a relation between similar and comparable quantities.
Example: Consider a sinusoidal oscillation which begins at time $t = 0$ with a phase δ (positive or negative). Then instead of $A\sin(\omega t + \delta)$, we could write (in the complex representation):

$$\zeta = A\,e^{i\delta} \cdot e^{i\omega t}\,. \tag{25.31}$$

The product $A\,e^{i\delta} = A'$ is called the *complex amplitude*. It contains *two* parameters of the oscillation, namely both the real amplitude and also the phase angle δ. The ratio of two complex amplitudes

$$\frac{A_1'}{A_2'} = \frac{A_1}{A_2} \cdot e^{i(\delta_1 - \delta_2)} = \varrho\, e^{i\delta} \tag{25.32}$$

contains both the ratio $\varrho = A_1/A_2$ of the real amplitudes as well as the phase difference δ between them. Here, ϱ is the magnitude and δ the phase angle of the complex number $\varrho\, e^{i\delta}$.

In a material with extinction, the waves are damped exponentially. After a distance z, the power has decreased to the fraction e^{-Kz} of its original value, and the amplitude has thus decreased[C25.12] to the fraction $e^{-Kz/2}$. If the extinction constant K is replaced by the extinction *coefficient* k using the relation

C25.12. The power is proportional to the square of the amplitude (see Comment C12.4).

$$K = \frac{4\pi k}{\lambda} \tag{25.3}$$

(λ: wavelength in vacuum), then we obtain for the momentary value at the location z and at time t

$$E_x = A \cdot e^{-2\pi kz/\lambda} \cdot e^{i\omega(t - zn/c)}\,. \tag{25.33}$$

The transition from Eq. (25.26) (a wave without extinction) to Eq. (25.33) (a wave with extinction) can be carried out formally in a different way: We need only replace the refractive index n in Eq. (25.26) by a complex quantity, namely the *complex index of refraction*:

$$n' = n - ik\,. \tag{25.34}$$

It contains *two* numerical quantities, both the refractive index n and the extinction coefficient k. Making use of the complex index of refraction, we can go directly from Eq. (25.26) to Eq. (25.33).

This result is important. It can be used to compute the influence of extinction on the propagation of a wave from a simple rule: We start with the formulas derived for a wave without extinction and replace the refractive index n there by the complex index of refraction $n' = n - ik$. It performs excellent service as a formal computational quantity and is indispensable in any treatment of the extinction of waves.

25.11 BEER's Formula for Perpendicular Reflection by Strongly-Absorbing Materials

We have already presented the experimental facts in Sect. 25.6. Our quantitative treatment is based on an extension of FRESNEL's formulas. In addition to the refractive index n, we must also take the extinction coefficient k into account. This is accomplished by applying the general rule introduced above: We replace the real refractive index n by the complex refractive index $n' = n - ik$.

In the special case of perpendicular incidence, we found for the reflection

$$\frac{E_\mathrm{r}}{E_\mathrm{i}} = -\frac{n-1}{n+1}\,. \qquad (25.15)$$

Inserting the complex refractive index, we obtain the ratio of two complex amplitudes:

$$\frac{E'_\mathrm{r}}{E'_\mathrm{i}} = -\frac{n-ik-1}{n-ik+1} = \varrho\, e^{i\delta_r}\,. \qquad (25.35)$$

Here, the magnitude ϱ (see Eq. (25.32) in Sect. 25.10) is the ratio of the real amplitudes, that is $\varrho = E_\mathrm{r}/E_\mathrm{i}$; and δ_r is the phase angle between $\boldsymbol{E}_\mathrm{r}$ and $\boldsymbol{E}_\mathrm{i}$, i.e. between the reflected and the incident amplitudes. Both are to be computed according to the rules in Sect. 25.10. We begin with the calculation of the

$$\text{Reflectivity } R = \varrho^2 = \left|\frac{E_\mathrm{r}}{E_\mathrm{i}}\right|^2 .$$

For this calculation, we multiply the complex number in Eq. (25.35) by its complex conjugate, thus

$$R = \frac{(n-ik-1)(n+ik-1)}{(n-ik+1)(n+ik+1)} \qquad (25.36)$$

or

$$R = \left|\frac{E_\mathrm{r}}{E_\mathrm{i}}\right|^2 = \frac{(n-1)^2+k^2}{(n+1)^2+k^2}\,. \qquad (25.37)$$

This is the often-used formula of AUGUST BEER (1854). For every value of the reflectivity R, there are many pairs of values of the optical constants n and k that satisfy Eq. (25.37). The set of all these pairs forms circles, as is shown in Fig. 25.15 for values of R between 20 and 80 %.

In metals, the summand k^2 in the numerator and denominator of BEER's formula (25.37) is generally predominant. Then R is comparable to 1. A large fraction of the incident radiant power is reflected. In the example in Fig. 25.6 c, this fraction was $(E_\mathrm{r}/E_\mathrm{i})^2 \approx 0.78^2 \approx 60\,\%$. Silver can reflect more than 95 % in the visible range. In the longer-wavelength infrared range, all metals have a reflectivity of $R \approx 100\,\%$ (cf. Fig. 27.8).

To calculate the phase difference, we put Eq. (25.35) in the form $a + ib$. To do this, we multiply the numerator and the denominator by the complex conjugate of the denominator, that is

$$\varrho\, e^{i\delta_\mathrm{r}} = -\frac{n-ik-1}{n-ik+1}\cdot\frac{n+ik+1}{n+ik+1} = \frac{1-n^2-k^2+i2k}{n^2+2n+1+k^2} \qquad (25.38)$$

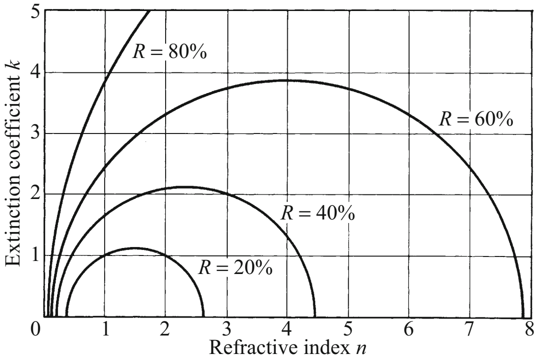


Figure 25.15 A graphical representation of Beer's formula shows pairs of values of n and k which give the same value of the reflectivity R at perpendicular incidence. The center of the circles is at $n = (1+R)/(1-R)$, and their radii r are given by $r^2 = 4R/(1-R)^2$

or

$$((n+1)^2 + k^2) \cdot \varrho\, e^{i\delta_\mathrm{r}} = \underbrace{1 - n^2 - k^2}_{\text{Real part}} + \underbrace{i2k}_{\text{Imaginary part}} .$$

Then we make use of Eq. (25.28),

$$\tan\delta_\mathrm{r} = \frac{\text{Imaginary part}}{\text{Real part}} \text{ of the complex quantity}, \tag{25.39}$$

and obtain for the phase angle between the reflected and the incident amplitudes:

$$\tan\delta_r = \frac{2k}{1 - n^2 - k^2} . \tag{25.40}$$

In a similar manner, we could start from the Fresnel formula (25.14) and calculate the ratio of the transmitted amplitude $\boldsymbol{E}_\mathrm{t}$ and the incident amplitude $\boldsymbol{E}_\mathrm{i}$, and likewise the phase angle δ_t between the amplitudes. For perpendicular incidence, we obtain

$$\left|\frac{E_\mathrm{t}}{E_\mathrm{i}}\right|^2 = \frac{4}{(n+1)^2 + k^2} \tag{25.41}$$

and

$$\tan\delta_\mathrm{t} = \frac{k}{n+1} . \tag{25.42}$$

In Fig. 25.9, we illustrated Fresnel's formula for perpendicular incidence and weak reflection with a 'snapshot' image, for the numerical example of $n = 2$. Analogously, Fig. 25.16 shows 'snapshot' images to elucidate Eqns. (25.37) through (25.42); at the left, for $n = 2$ and $k = 4$, and at the right, for $n = 2$ and $k = 0.1$.

Figure 25.16 (right) is hardly distinguishable from Fig. 25.9 (top). This means that an absorption coefficient of $k = 0.1$ plays practically

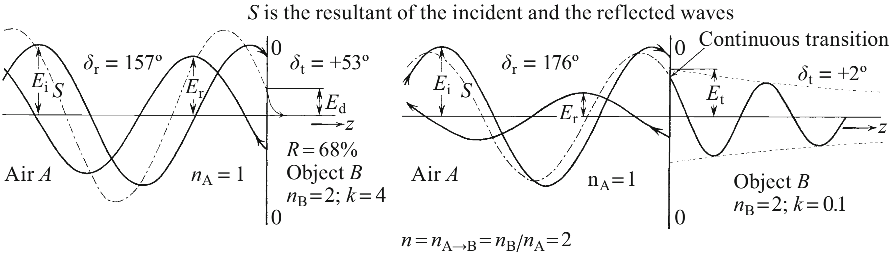


Figure 25.16 'Snapshot' images of waves, continuing those in Fig. 25.9, to illustrate the application of Eqns. (25.37) through (25.42). At the boundary between air and the object B, at each moment the magnitude of the light vector of the transmitted light is equal to the sum of the light vectors of the incident and the reflected light. The left-hand image represents for example the reflection of red light from a platinum surface. The right-hand image exaggerates somewhat the situation for dye solutions of very high concentrations.

no role in reflection. $k = 0.1$ (or, more precisely, $k = 0.08$) means that $w = \lambda$, i.e. the mean penetration depth of the light is equal to its wavelength (in vacuum). $w = \lambda$ was introduced in Sect. 25.5 as the boundary between strong and weak absorption. That definition finds its justification here.

If, for a strongly-absorbing material, we have measured two of the three quantities R, n and $k = K\lambda/4\pi$, then we can use BEER's formula (25.37) to calculate the third. We could however also measure R and δ_r, and then combine Eqns. (25.37) and (25.40) in order to obtain k and n.

25.12 Light Absorption by Strongly-Absorbing Materials at Oblique Incidence

In Sect. 25.11, we have discussed the reflection of light with strong extinction at perpendicular incidence ($\alpha = 0$) rather thoroughly. The significance of the equations derived there extends far beyond the field of optics. These equations also play an important role in acoustics and electrical technology. They indeed contain only two formally-defined materials constants, the refractive index n and the absorption coefficient k, independently of any considerations of the exact nature of the waves.

When the light is incident at an oblique angle ($\alpha > 0$), the situation becomes more complicated. If we insert the complex index of refraction into the law of refraction, we obtain a complex angle of refraction. It contains two pieces of information: First, the positions of surfaces of equal phase, and second, the positions of surfaces of equal amplitude. Figure 25.17 serves to illustrate this. In the figure,

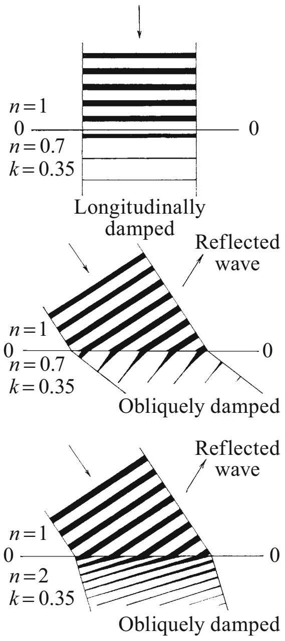


Figure 25.17 The various forms of spatial damping of travelling waves (the thickness of the lines indicates the amplitude of the waves)

the wave crests are marked by broad dark lines. Their thickness is supposed to indicate the amplitudes. In the first two images, the refractive index of the material below the boundary $O - O$ is presumed to be smaller than that of the material above the boundary.

In Fig. 25.17, top, $\alpha = 0$, so that the light is incident perpendicular to the boundary $O{-}O$. The lines of constant phase (the wave crests) and the lines of constant amplitude (equal thickness of the lines drawn) coincide: We have longitudinal damping.

In the center image of Fig. 25.17, α is about 33°. Now, the wave crests below the boundary no longer coincide with the lines of constant amplitude, i.e. in the figure with lines of constant thickness. The wave is “inhomogeneous” and “obliquely damped”.

In Fig. 25.17, bottom, the refractive index of the material below the boundary is larger than it is above. Here, again, we see an oblique damping.

Experimentally, this oblique damping makes itself known in an unpleasant manner: The ratio $\sin\alpha/\sin\beta$ measured with prisms is no longer constant; it depends upon the angle of incidence (Fig. 25.18) and can increase by more than a factor of two with increasing α, for example in the case of Cu.

In spite of these complications, the case of oblique incidence on strongly-absorbing materials can be treated just like the case of perpendicular incidence. We once again start with FRESNEL’s formulas for weak absorption, that is with Eqns. (25.11) and (25.12). Once

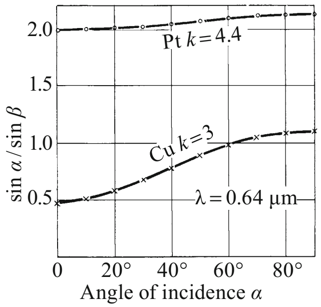


Figure 25.18 In materials with strong absorption, the ratio $\sin\alpha/\sin\beta$ depends on the angle of incidence α (this was measured by D. SHEA using very thin metal prisms)

again, we replace the real refractive index n by a complex index of refraction, which also takes absorption into account:

$$n' = n - ik\,. \tag{25.34}$$

Unfortunately, the ensuing computations are rather extensive and complicated, if carried out in rigorous form. For this reason, we limit the problem and ask only, "How can we determine the optical constants n and k from *reflection* measurements with *oblique* incidence of the light"?

For the special case that α is equal to the principal angle of incidence $\varPhi$ (Sect. 25.6, Fig. 25.6c), we have, according to CAUCHY's formulas[C25.13],

$$k = n\tan 2\varPsi \tag{25.43}$$

$$n = \sin\varPhi\tan\varPhi\cos 2\varPsi\,, \tag{25.44}$$

where $\varPsi$ is defined by

$$\tan\varPsi = \left(\frac{E_{\mathrm{r}\parallel}}{E_{\mathrm{r}\perp}}\right)_{\alpha=\varPhi}\,. \tag{25.45}$$

We thus have two equations for the determination of the two optical constants n and k. The measured quantities are the principal angle of incidence $\varPhi$ and $\tan\varPsi$, that is the ratio of the two reflected amplitudes at the principal angle of incidence (Eq. (25.45) and Fig. 25.6c).

The two equations (25.43) and (25.44) are rather important in measurement technology. They were published already in 1849 by A.L. CAUCHY. They should thus not be considered to be results of MAXWELL's theory, contrary to popular belief.

25.13 Conclusion: Pictures Used in Physical Descriptions

The quantitative treatment of "strong" light absorption, when $w < \lambda$, is not a pleasing chapter. One has to carry out numerous calculations and nevertheless, for oblique incidence, only approximate solutions lead to formulas of acceptable simplicity.

C25.13. POHL gave a derivation of CAUCHY's formulas in the 13th edition of "*Optik und Atomphysik*", Chap. 11. See also PAUL DRUDE, *Annalen der Physik* **271** (1888), p. 508–523. A modern derivation in English is given in Born and Wolf's book, Sect. 13.2 (cf. Comment C25.11.).

"Even beginning physics students associate optical measurements with a high degree of precision; they are aware that refractive indices, wavelengths, etc. are measured to many significant figures. In the case of strong absorption, this precision goes out the window".

But another aspect is still worse. Even beginning physics students associate optical measurements with a high degree of precision; they are aware that refractive indices, wavelengths, etc. are known to many significant figures. In the case of strong absorption, this precision goes out the window. Being able to reproduce measurements of n and k to within a few percent has to be considered satisfactory. The reason is clear: With strong absorption, all of the processes take place within very thin surface layers of the absorbing material; the main contributions are made by layers less than 10^{-4} mm thick. These layers, in contrast to the bulk of the material, are unprotected against all kinds of external influences. Their structures are not stable over time; they depend on the history of the material and on the presence of impurity molecules near the surface. The situation is similar to that of external friction between two solid objects in close contact.

No surface layer shows the same properties as the bulk of the material. For example, we can place a glass block with a carefully polished surface into a liquid with exactly the same refractive index (for the light being used). But we can always see the surface layer owing to reflections of up to some tenths of a percent of the incident light. The refractive index of the surface layer is always somewhat different from that of the bulk glass. According to Lord RAYLEIGH (1937)[C25.14], the thickness of the layer that is influenced by handling is about $3 \cdot 10^{-6}$ cm (0.03 μm), and the increase in its refractive index can be up to 10 %.

C25.14. *Proceedings of the Royal Society* (London), Series A, Vol. **160**, pp. 507–526 (1937).

This fact is particularly noticeable in the filters fabricated by the method of CHRISTIANSEN. These filters consist of a layer at least 1 cm thick of fine, carefully purified glass powder in a mixture of benzene and carbon disulfide. At the correct mixing ratio, the dispersion curve of the glass and that of the liquid can be made to intersect. Then the glass and the liquid have practically the same refractive index over a narrow frequency range; for the transition from glass → liquid, $n = 1$. Light within this range is transmitted without attenuation, while everything outside the range is deflected off to the sides by diffuse reflection. This is however only approximately realized, because the grains of glass powder have no unified value of their refractive index near their surfaces.

All physical descriptions make use of simplifying pictures which in the end are useful only as approximations. The same empirical facts may be encompassed by different pictures. The simplifications must be kept to that minimum which is still compatible with the intended purpose of the pictures. An example is more instructive than wordy explanations:

In drawings, for example in a sketch of a lens, we indicate the boundaries of a body by a surface. A surface is a simplified picture: In reality, we are dealing with an inhomogeneous transition layer of finite thickness. If a surface is described as *planar*, we are again using a simplifying picture.

Physically, a fresh liquid surface, for example of water, exhibits the least irregularities. But every liquid has a vapor pressure, e.g. for water at room temperature, it is 24 hPa. Therefore, at the boundary between the liquid and its vapor, there is a statistical equilibrium between evaporating and condensing molecules. Per square centimeter and second, around 10^{22} molecules make the transition from the liquid into the vapor and *vice versa*. In a square centimeter of the surface, however, there is room for only 10^{15} molecules. Each individual molecule can thus remain at the surface for only about 10^{-7} s; then it again flies off the surface with a velocity of around 700 m/s. This clamoring, swarming throng is the best approximation that physicists can use to approach the ideal picture of a surface as formulated by mathematics!

"This clamoring, swarming throng is the best approximation that physicists can use to approach the ideal picture of a surface as formulated by mathematics!"

All pictures and words are contingent on their time. They must be adjusted in the course of years to the continual extension of our experimental knowledge.

"All pictures and words are contingent on their time. They must be adjusted in the course of years to the continual extension of our experimental knowledge".

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_25) contains supplementary material, which is available to authorized users.

Scattering 26

26.1 Preliminary Remark

In the preceding chapters, we have described *quantitatively* the propagation of radiation from its source to a receiver using *two* quantities, usually the refractive index n and the extinction coefficient k. Within a *qualitative* description, the phenomena related to diffuse reflection and scattering were also considered. Both play an important role in optics. They lead us to the concept of a light beam and its graphical representation using 'light rays' drawn as straight lines. Diffuse reflection and scattering allow objects that themselves do not emit light to become visible as "secondary emitters". The treatment of some important diffraction and interference phenomena is based upon them. Scattering allows us to identify polarized light through its asymmetry (Fig. 24.4).

These examples however by no means exhaust the significance of scattering. Scattering leads to a whole series of other important insights, for example in connection with refraction and dispersion (Chap. 27). This is why we want to treat the topic of scattering more comprehensively in this chapter.

26.2 The Basic Ideas Underlying the Quantitative Treatment of Scattering

The fundamental aspects of scattering have already been illustrated by demonstration experiments in earlier chapters. Their qualitative interpretation makes use of the analogy to water waves: An obstacle which is small compared to the wavelength, e.g. a rod, is encountered by a wave train. The obstacle then becomes the source of a new, "secondary" wave train which propagates in all directions away from it (Vol. 1, Fig. 12.17).

The obstacle is presumed to be rigid and immobile. This is however only a special case. In general, the obstacle will be an object which can itself vibrate (an *oscillator*), and it can be excited to *forced vibrations* as a *resonator* by the oncoming waves. Forced vibrations of harmonic oscillators (sinusoidal oscillations) were treated in depth

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_26

in Vol. 1 (Sect. 11.10)[1]. Here, we review briefly the most important aspects and supplement them by quantitative data.

A force $F = F_0 \cos(2\pi\nu t)$ which acts periodically on a harmonic oscillator (e.g. a sinusoidally-vibrating spring pendulum with a mass m, see Vol. 1, Fig. 4.13) produces forced vibrations, as shown in Vol. 1, Fig. 11.42b for torsional oscillations. In a steady state, their amplitude depends on F_0 and on the frequency ν, and also on the eigenfrequency ν_0 of the free oscillator, and on its damping, expressed as the logarithmic decrement Λ. Quantitatively, we found for the amplitude (cf. Comment C11.8 and **Exercise 26.1**):

$$l_0 = \frac{1}{4\pi^2} \frac{F_0/m}{\sqrt{(\nu_0^2 - \nu^2)^2 + \left(\frac{\Lambda}{\pi}\right)^2 \cdot \nu_0^2 \nu^2}} . \tag{26.1}$$

The oscillator vibrates at the frequency ν, but with a phase shift of φ:

$$l(t) = l_0 \cos(2\pi\nu t - \varphi) , \tag{26.2}$$

where

$$\tan\varphi = \frac{\Lambda}{\pi} \cdot \frac{\nu_0 \nu}{\nu_0^2 - \nu^2} . \tag{26.3}$$

These forced vibrations, for their part, cause the *emission* of secondary waves. In the case of light scattering, we must describe the mechanism of this emission quantitatively. We undertake this task in the following section.

26.3 The Radiation from Oscillating Dipoles. PURCELL's Experiment

The analogous behavior of electrical waves and light waves has already been pointed out in Sect. 12.8. Here, we extend that discussion by a comparison which forms the basis for a treatment of scattering.

Conveniently-operated transmitters for linearly-polarized electromagnetic waves of short wavelength are readily available today. We make use of one of them (Fig. 26.1). Its essential component is easily recognized, namely the short antenna S. High-frequency alternating current flows in it. The devices required to generate this current and the technical accessories (electronic components, etc.) are contained in the shielded box K. The electric field produced by the transmitter lies in planes which contain the long axis of the antenna S.

[1] The analogous forced oscillation of an RLC circuit (tank circuit) was also described in Sect. 11.7.

Figure 26.1 Left: A transmitter or source dipole for emitting undamped waves ($\lambda \approx 10\,\mathrm{cm}$). Right: A non-tuned receiver dipole with a rectifier and a galvanometer as detector

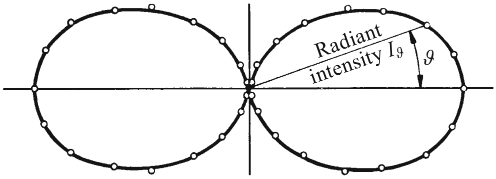
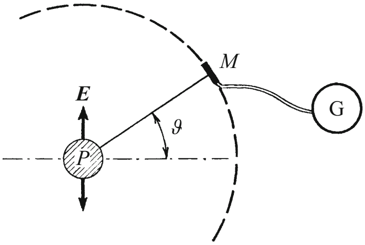


Figure 26.2 The graph shows how the radiant intensity I_ϑ of the waves emitted by the source depends on the angle ϑ between the direction of propagation of the waves and a plane perpendicular to the long axis of the source. For the measurement, the receiver (antenna E) is perpendicular to the direction of propagation of the waves, while the angle is varied by tilting the source antenna S. At $\vartheta = 0$, the source and the receiver antennas are parallel.

As receiver, we use a short antenna E (as described already in Sect. 12.6). At its center, it contains a rectifier (diode), which produces a direct current that is measured by the ammeter (galvanometer G). With this setup, we measure the radiant intensity I_ϑ (Eq. (19.2)) of the linearly-polarized radiation as a function of the angle ϑ. The result is plotted in Fig. 26.2. This corresponds to Fig. 12.24.

Now, we show a corresponding experiment in optics. In Fig. 24.4, we produced linearly-polarized light by scattering. We repeat that experiment in quantitative form here. In Fig. 26.3, the shaded circle P is the cross-section of the primary light beam within a cloudy medium. Its plane of oscillation is marked by a double arrow $\boldsymbol{E}$. Along the large dashed circle, we can slide a radiometer M around the beam P which forms the center of the circle. We measure the intensity of the scattered radiation (i.e. the deflection of the ammeter) as a function of the angle ϑ. The result can be seen in Fig. 26.4, as the solid curve. The similarity between Figs. 26.4 and 26.2 is evident. In both cases, for the radiant intensity I_ϑ in the direction ϑ, we find to a good

Figure 26.3 The measurement of the scattered radiation as a function of the angle. At P, the primary beam of linearly-polarized light is incident perpendicular to the plane of the page.



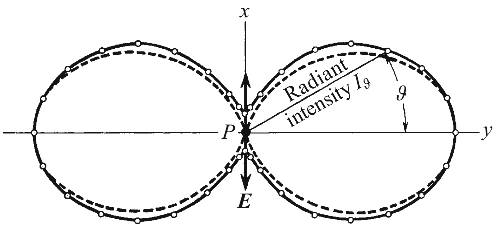


Figure 26.4 The scattering of polarized light by spherical dielectric particles. The primary light beam is perpendicular to the plane of the page at the point P, and $\boldsymbol{E}$ indicates its plane of oscillation. The length of the radius corresponds to the radiant intensity (or to the deflection of the radiometer M in Fig. 26.3). The figure can be thought of as rotationally symmetric around the double arrow $\boldsymbol{E}$ as central axis (see also Fig. 26.9).

approximation

$$I_\vartheta = \text{const} \cdot \cos^2 \vartheta \,. \tag{26.4}$$

This relation is shown by the dashed curve in Fig. 26.9.

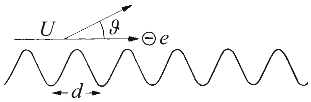
This analogous behavior leads to the following conclusions: In the optical experiment, the incident polarized light converts the suspended particles in the cloudy liquid into *sources*, which radiate as dipole antennas. The light can excite the suspended particles because it consists itself of electromagnetic waves. Their electric fields can produce periodically alternating electric dipole moments in the suspended particles; or, put briefly, excite them to *forced electrical oscillations*.

E.M. Purcell described an experiment in which *visible* light is produced as dipole radiation.[C26.1] This demonstration is the electrical analogue of the acoustical experiment with which Thomas Young in 1801 explained the action of a grating (Sect. 22.4, small print at the end). Its principle: In Fig. 26.5, an electron passes with the velocity u closely above a corrugated metal sheet. Its negative charge, together with the positive influence charge in the metal (mirror charge), form a dipole. The spacing of the two charges and thus the dipole moment are varied periodically with the period $T = d/u$. This corresponds to a frequency of $\nu = u/d$. In a direction ϑ, as a result of the Doppler effect (Sect. 23.5), the frequency $\nu' = \nu / \left(1 - \frac{u}{c} \cos\vartheta\right)$ will be observed, corresponding to the wavelength

$$\lambda' = d \left(\frac{c}{u} - \cos\vartheta \right) .$$

C26.1. See S.J. Smith and E.M. Purcell, *Physical Review* **92**, 1069 (1953).

Figure 26.5 The production of *visible* dipole radiation



Example: An optical grating with $d \approx 1.7\,\mu\text{m}$ serves as a "corrugated sheet". Close to its surface and at right angles to its grooves, there is a thin electron beam (diameter ≈ 0.15 mm, accelerating voltage $U = 3 \cdot 10^5$ V, $I = 5 \cdot 10^{-4}$ A, $u \approx c$). One can see its path as a colored streak whose hue changes with ϑ.

26.4 Quantitative Treatment of Dipole Radiation

In an electric field, every object becomes an electric dipole: Every conductor through influence (Fig. 2.25 b), and every insulator through the "polarization of the dielectric". This can occur in two different ways: First, through the action of influence on the individual molecules (Fig. 2.56); and second, through a parallel alignment of the dipoles of "polar" molecules that are already present without the field, but are randomly oriented due to thermal motions. These are molecules with permanent electric dipole moments, e.g. H_2O and HCl (Sect. 13.10). *These polar molecules will initially be left out of our considerations*. They will be treated later in Sect. 27.16.

An oscillating dipole is the archetype of an electromagnetic wave source (HEINRICH HERTZ, 1887). In the simplest case, its electric dipole moment p varies sinusoidally, so that we have:

$$p = p_0 \sin \omega t\,. \tag{26.5}$$

Let the amplitude of the dipole moment be $p_0 = Q\,l_0$. Then at a large distance r (i.e. $r \gg l_0$, the length of the dipole), the radiant intensity from the dipole in the direction ϑ is given by

$$I_\vartheta = \frac{c\pi^2}{2\varepsilon_0} \cdot \frac{p_0^2}{\lambda^4} \cos^2 \vartheta \tag{26.6}$$

(Units: watt/steradian (Eq. (19.2)); c is the velocity of light, and ε_0 is the electric field constant = $8.86 \cdot 10^{-12}$ A s/V m).

How Eq. (26.6) comes about is easy to understand qualitatively:C26.2 Assuming that the dipole is undergoing forced oscillations at the circular frequency $\omega = 2\pi\nu$, then the electric field that it emits is produced by an *induction* process, so that its amplitude E_0 is proportional to the time derivative of the *electric current*. Furthermore, the current flowing in the oscillating dipole is $\sim \mathrm{d}p/\mathrm{d}t$. Because of this second-order differentiation, the amplitude E_0 of the emitted field is $\sim -\omega^2 p_0$, and its power is thus $\sim \omega^4 p_0^2 \sim p_0^2/\lambda^4$. (The minus sign before $\omega^2 p_0$ means that there is a phase difference of 180° between the field emitted and the dipole moment).

Integration of the radiant intensity in Eq. (26.6) over the solid angle $\mathrm{d}\Omega$ (that is over ϑ and φ, total solid angle $\Omega = 4\pi$; see Fig. 19.1) yields the total average power emitted by the dipole at the frequency ν:

$$\overline{\dot{W}} = \frac{4c\pi^3}{3\varepsilon_0} \cdot \frac{p_0^2}{\lambda^4} = \frac{1}{12\pi\varepsilon_0 c^3} \cdot \omega^4 p_0^2 = \frac{4\pi^3}{3\varepsilon_0 c^3} \cdot \nu^4 p_0^2\,. \tag{26.7}$$

C26.2. A quantitative derivation can be found for example in F. Hund, "*Theoretische Physik*", Vol. 2, Sect. 61 (B.G. Teubner, Stuttgart, 1957) or in P. Lorrain, D.R. Corson, and F. Lorrain, "*Fundamentals of Electromagnetic Phenomena*", Chap. 25 (W.H. Freeman, New York, 2000).
The integration over solid angles leading from Eq. (26.6) to Eq. (26.7) can be found in standard textbooks on integral calculus. Note that the angle ϑ as defined here (Fig. 26.9) is the complement of the usual polar angle ϑ in spherical coordinates. φ is the azimuthal angle around the z axis.

26.5 The Wavelength Dependence of RAYLEIGH Scattering

Now, we have all the prerequisites for a quantitative treatment of scattering. RAYLEIGH scattering is characterized by the following assumptions: The scattering particles are spherical and their diameters are small compared to the wavelength of the light that is scattered. They are transparent in the visible spectral range and their absorption begins only in the ultraviolet, and thus at higher frequencies. Furthermore, their arrangement in space allows no fixed phase relations to be established among the rays of secondary radiation from the individual scattering particles.$^{C26.3}$ For this reason, the average spacing of the particles is supposed to be larger than the wavelength, and their arrangement is statistically disordered. In Fig. 24.4, we illustrate these conditions by using fine particles of a weakly-absorbing material suspended in water. Scattering by these particles leads to extinction of the primary light beam. Its measurement, as described for example in Sect. 25.2, yields an extinction constant K proportional to the number density N_{V} of the scatterers. Thus, at a given wavelength, the quotient K/N_{V}, called the scattering or interaction cross-section (Sect. 25.4), is constant.

C26.3. The important role played by the phase relations between the sources of scattered radiation is emphasized by the footnote near the end of this section. These relations lead to a reduction of the overall scattered radiant intensity; that is, they increase the fraction of the light which continues along the path of the incident light beam.

We begin by following RAYLEIGH's method of calculating the wavelength dependence of the extinction constant K due to scattering alone. The light beam is assumed to be *collimated*, i.e. it has parallel bounds; then in a given section of the beam of length Δx and cross-sectional area A, there will be $N_{\mathrm{V}} A \Delta x$ scattering particles. They give rise to an extinction constant of:

$$K = \frac{\Delta\overline{\dot{W}}}{\overline{\dot{W}}_{\mathrm{p}}}\frac{1}{\Delta x}\,. \qquad \text{(Defining equation (25.1))}$$

Here, $\Delta\overline{\dot{W}}$ indicates the average radiant power (or the radiant intensity) of the secondary radiation, and

$$\overline{\dot{W}}_{\mathrm{p}} = \frac{\varepsilon_0}{2}E_0^2 c \cdot A \qquad (26.8)$$

(ε_0 is the electric field constant, E_0 the amplitude of the electric field of the light, and c is the velocity of light)

is the radiant power in the primary beam of light that passes through the area A.$^{C26.4}$ $\Delta\overline{\dot{W}}$ is composed of the sum of the intensities given by Eq. (26.7) from all of the scattering particles within the volume $A\Delta x$:

C26.4. The derivation is given in Comment C27.9.

$$\Delta\overline{\dot{W}} = N_{\mathrm{V}} A \Delta x \frac{4c\pi^3}{3\varepsilon_0} \cdot \frac{p_0^2}{\lambda^4}\,. \qquad (26.9)$$

In this expression, $p_0 = Q\,l_0$ is the amplitude of the dipole moment of a scatterer which is induced by the field E_0 of the primary light that excites the scattering.

Combining Eqns. (26.8) and (26.9) with the definition (25.1) yields the extinction constant due to scattering:

$$K = N_\mathrm{V} \frac{8\pi^3}{3\varepsilon_0^2} \left(\frac{p_0}{E_0} \right)^2 \cdot \frac{1}{\lambda^4}\,. \qquad (26.10)$$

The relation between $p_0 = Q\,l_0$ and the field strength E_0 can be calculated quite generally from Eq. (26.1). The amplitude of the force is given by $F_0 = QE_0$. The scattering particles are small compared to the wavelength; therefore, considered as antennas, they have very high eigenfrequencies ν_0. Compared to these eigenfrequencies, we can neglect the frequency ν of the primary light in Eq. (26.1). Then the amplitude and thus also the polarizability $\alpha = Q\,l_0/E_0 = p_0/E_0$ are independent of ν, so that only constant quantities occur in Eq. (26.10) before $1/\lambda^4$, and we obtain

$$K = \text{const} \cdot \frac{1}{\lambda^4}\,. \qquad (26.11)$$

The extinction constant resulting from this so-called RAYLEIGH *scattering* is thus proportional to $1/\lambda^4$ (like the power emitted by the dipoles).

This important relation (26.11) is realized experimentally only as a limiting case. A good example is the scattering in a NaCl crystal with small additions of $SrCl_2$ (Sr^{++} ions/Na^+ ions $= 1 : 10^3$). These additional ions produce local lattice perturbations in the crystal. In reflected daylight, the crystals appear bluish, but they are reddish-yellow with transmitted light. Figure 26.6 shows measurements of the constant K due to scattering in the wavelength range between $\lambda = 0.2\,\mu$m and $\lambda = 1\,\mu$m. The coordinate axes are logarithmic. The measured data points lie on the solid straight line, which corresponds to $K = \text{const}/\lambda^{3.8}$. The dashed line would correspond to $K = \text{const}/\lambda^4$; thus, Eq. (26.11) is fulfilled to a good approximation, but not strictly verified. The approximation would in any case be sufficiently good to determine one of the two quantities N_V or p_0/E_0, that is the particle number density and their polarizability, if the other quantity is known.

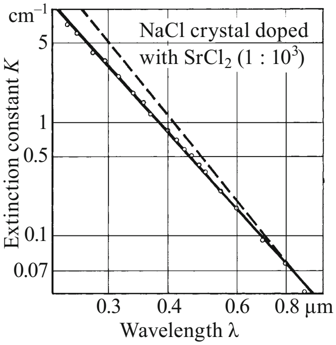
Qualitative examples for preferential scattering at shorter wavelengths are easily found.[C26.5] Water containing some drops of milk looks bluish. Thin skin over the dark background of veins near the surface also looks bluish, for example on the inside of the wrist. The most famous example is provided by the earth's atmosphere. It scatters the shorter-wavelength light in the visible spectrum; thus, the clear sky appears blue.

During the day, even when we are standing in shadow, we cannot see the stars; we are dazzled by the secondary light from scattering in the

C26.5. See for example **Video 16.1 ("Polarized light")** http://tiny.cc/5dggoy, or, in Vol. 1, **Video 10.4 ("Smoke rings")** http://tiny.cc/ocgvjy.

C26.6. In the wavelength range of these measurements, K varies by more than a factor of 200! Light scattering with a similar wavelength dependence was also observed in NaCl:Mn and KCl:Ca. The deviation from the expected wavelength dependence (Eq. (26.11)) was explained by the fact that the scattering particles have a diameter of ca. 150 nm (onset of MIE scattering). (K.G. Bansigir and E.E. Schneider, *Journal of Applied Physics*, Supplement to Vol. **33**, p. 383 (1962)).

Figure 26.6 The wavelength dependence of the extinction constant for RAYLEIGH scattering[C26.6]



upper atmosphere. The longer the path of light through the air, the greater is the loss by extinction due to scattering. As a result, we see the sun's disk near the horizon with a bearable brightness and colored yellow-orange to red.

In the clear, dust-free atmosphere, only the individual molecules act as scattering centers[2]. Therefore, from the measured extinction constant K of the atmosphere, we can determine the number density of the molecules. This is carried out as follows: Quantitatively, Eq. (26.10) holds. However, we now denote the dipole moment of a single molecule as p_0', that is $p_0' = Q\,l_0$. For its polarizability, we have $p_0'/E_0 = \alpha$ (cf. Sect. 13.9, Eq. (13.26)). Then Eq. (26.10) can be rewritten as:

$$K = N_\mathrm{V}\frac{8\pi^3}{3\varepsilon_0^2}\cdot\alpha^2\cdot\frac{1}{\lambda^4}\,. \tag{26.12}$$

With $\nu \ll \nu_0$, the polarizability α becomes independent of the frequency and has the same value as in a static field. The polarizability α of single molecules of a material is well known from electromagnetic theory; it was determined in Sect. 13.9 from the dielectric constant ε. For gases with $\varepsilon \approx 1$ (Eq. (13.27)), it was shown there that

$$\alpha = \frac{\varepsilon_0}{N_\mathrm{V}}(\varepsilon - 1)\,. \tag{26.13}$$

C26.7. See for example Max Born and Emil Wolf, "Principles of Optics" (4th ed., Pergamon Press 1970), Sect. 81; available online: See Comment C25.11. A brief introduction can also be found in F.S. Crawford, "Waves", *Berkeley Physics Course*, Vol. 3 (McGraw Hill, New York 1968), p. 559.

[2] The average spacing of the molecules is in fact small compared to the wavelength (near the earth's surface, it is about $3\cdot 10^{-9}$ m); but the large local thermal density fluctuations in gases act to eliminate phase relations between the secondary rays from individual molecules. This can be shown quantitatively.[C26.7] Liquids are less compressible than gases and vapors. Their thermal motions therefore produce much smaller statistically-distributed density fluctuations than in gases and vapors. As a result, light scattering by liquids is relatively weak. To demonstrate it clearly, we first have to remove all suspended particles from the liquid by distillation in vacuum. For demonstration experiments, benzene or diethyl ether are suitable; in both, light scattering can be observed using red-filter light. The local density fluctuations in solids are even smaller than in liquids. In a block of good-quality optical glass with polished faces, the scattering cone can still be readily observed. A similar block of crystalline quartz has to be heated to several hundred °C to make the scattering visible.

Combining Eqns. (26.12) and (26.13), we obtain

$$N_{\mathrm{V}} = \frac{8\pi^3}{3K} \cdot \frac{(\varepsilon - 1)^2}{\lambda^4} \,. \tag{26.14}$$

Observations (e.g. from the *Pic de Teneriffe*) have given a roughly constant value of the product $K \cdot \lambda^4 = 1.13 \cdot 10^{-30}\,\mathrm{m}^3$ at 0 °C and 1013 hPa between $\lambda = 320$ and 480 nm. Thus, for example, at $\lambda = 375\,\mathrm{nm}$, the extinction constant is $K = 5.7 \cdot 10^{-5}\,\mathrm{m}^{-1}$. This is an extraordinarily small value. It means that a reduction in radiant intensity to $1/e = 37\,\%$ occurs only after a path length of 18 km! The dielectric constant of the air is $\varepsilon = 1.00063$. With these numerical values, Eq. (26.14) gives

$$N_{\mathrm{V}} = 2.9 \cdot 10^{25}\,\mathrm{m}^{-3} \,.$$

The number density $N_{\mathrm{V,id}}$ of an ideal gas under these conditions of pressure and temperature is known (Vol. 1, Sect. 14.6); it is

$$N_{\mathrm{V,id}} = \frac{p}{kT} = 2.7 \cdot 10^{25}\,\mathrm{m}^{-3}$$

($p = 1.013 \cdot 10^5$ hPa, $T = 273$ K, k (BOLTZMANN's constant) $= 1.38 \cdot 10^{-23}$ W s/K).

The good agreement of the two number densities verifies RAYLEIGH's theory, and with it, the perhaps initially unclear footnote on the previous page.

Finally, we investigate the connection between RAYLEIGH scattering and compressibility. The isothermal compressibility (Vol. 1, Sect. 14.9) is:

$$\kappa = \frac{\mathrm{d}V}{\mathrm{d}p} \cdot \frac{1}{V} \,.$$

For an ideal gas ($pV = NkT$), the magnitude of κ is

$$\kappa = \frac{1}{p} \,.$$

With this, we obtain from Eq. (26.14) for the extinction constant K

$$K = \frac{8\pi^3}{3} \cdot \frac{(\varepsilon - 1)^2}{\lambda^4} \cdot \kappa k T$$

In *ideal* gases, the product $\kappa T = T/p$ is independent of T, and therefore, so is K. In *real* gases, the product κT in the neighborhood of the critical point becomes very large (Vol. 1, Sect. 15.1), and thus so does K; that is, the light scattering is strong ("critical scattering"). This illustrates the importance of local density fluctuations near the critical point (compare the footnote on the previous page).

26.6 The Extinction of X-rays by Scattering

The extinction of X-rays by scattering depends in general in a complex manner on their wavelength and on the molar mass $M_\mathrm{n} = M/n$ (M is the mass and n the amount of substance) of the irradiated material. But here, also, a special case of scattering which is characterized by great simplicity has been found. It is illustrated in Fig. 26.7.

For this scattering by materials with a small molar mass M_n, there is a wavelength range in which the extinction constant relative to the density, K/ϱ, is independent of the molar mass of the scatterer and of its chemical bonding, where it takes on the practically constant value

$$\frac{K}{\varrho} = 0.02\,\frac{\mathrm{m}^2}{\mathrm{kg}}\,. \tag{26.15}$$

The scattering in this characteristic wavelength range has led us to two important physical perceptions: First, we can see that the number Z of electrons in atoms with moderate molar masses is close to half as large as the number A of nucleons in their atomic nuclei (Sect. 26.7). Second, it has provided the possibility of producing and investigating linearly-polarized X-rays (Sect. 26.8).

26.7 The Number of Scattering Electrons in Light Atoms

Scattering of short-wavelength X-rays is independent of the bonding of the atoms in molecules or crystals. This is because for X-rays,

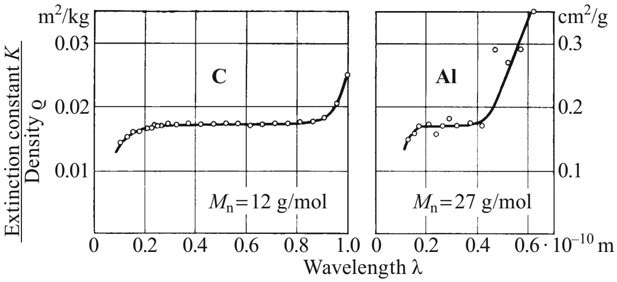


Figure 26.7 The influence of the wavelength on the scattering of X-rays by light atoms. On the ordinate, the extinction constant relative to the density ϱ, K/ϱ, is plotted against the wavelength λ as abscissa. Here, K is the extinction constant from scattering alone. (After measurements by C.W. HEWLETT; the portion of the total extinction constant due to absorption was subtracted. M_n is the molar mass of the scatterer).

only the electrons in the inner shells of the atoms act as scattering centers. If an atom has Z electrons, then the number density of its electrons is

$$N_\mathrm{V} = Z \cdot \frac{N_\mathrm{A}\varrho}{M_\mathrm{n}} \tag{26.16}$$

(N_A is the AVOGADRO constant $= 6.022 \cdot 10^{23}\,\mathrm{mol}^{-1}$, and M_n is the molar mass $= M/n$).

The atomic electrons are somehow able to oscillate around the positive charge which binds them to the nucleus of the atom. The oscillating electric field of the incident radiation excites the electrons to forced oscillations around their average rest positions. The positive charge remains at rest, due to the large mass of the atomic nucleus. The diameter of the electrons is small compared to the wavelength of the radiation, and their distribution within the atom is on average statistically disordered. Thus the conditions agree with those for RAYLEIGH scattering. For the extinction constant due to scattering, we can again use the equation

$$K = N_\mathrm{V}\frac{8\pi^3}{3\varepsilon_0^2} \cdot \alpha^2 \cdot \frac{1}{\lambda^4}\,. \tag{26.12}$$

Now, however, there is an essential difference: The eigenfrequencies ν_0 of the bound electrons in light atoms are small compared to the frequency ν of X-rays. As a result, their polarizability α is no longer constant, but rather it increases proportionally to λ^2. Therefore, K becomes independent of λ in Eq. (26.12). Derivation:

We again put $F_0 = eE_0$ into Eq. (26.1) (e is the electronic charge), but this time, we neglect ν_0 as small compared to ν. We thus obtain for the amplitude of the oscillations of the electrons

$$l_0 = \frac{1}{4\pi^2} \cdot \frac{e}{m\nu^2} \cdot E_0\,,$$

or, after multiplication by the charge e,

$$\frac{e\,l_0}{E_0} = \alpha = \frac{1}{4\pi^2} \cdot \frac{e^2}{m\nu^2} = \frac{e^2}{m} \cdot \frac{\lambda^2}{4\pi^2c^2}\,. \tag{26.17}$$

On inserting this expression for α into Eq. (26.12), the wavelength λ drops out. The remaining expression is

$$K = N_\mathrm{V}\frac{e^4}{6\pi\varepsilon_0^2 m^2 c^4} \tag{26.18}$$

(K is the extinction constant due to scattering, N_V the number density of the electrons, e the electronic charge $= -1.6 \cdot 10^{-19}$ A s, m is the electron's mass $= 9.1 \cdot 10^{-31}$ kg, ε_0 is the electric field constant $= 8.86 \cdot 10^{-12}$ A s/V m, and c the velocity of light $= 3 \cdot 10^8$ m/s).

Once more, in words: In the spectral range considered, the extinction constant K due to scattering of X-rays is independent of their wavelength; λ does not occur in Eq. (26.18). The equation contains only constants, apart from the number density of the electrons, N_V. Evaluating the constants yields for *one* electron

$$\frac{K}{N_\mathrm{V}} = 6.6 \cdot 10^{-29}\,\mathrm{m}^2\,. \tag{26.19}$$

With Z electrons per atom, we find from this, also using Eq. (26.16):

$$\frac{K}{\varrho} = 6.6 \cdot 10^{-29}\,\mathrm{m}^2 \cdot Z \cdot \frac{N_\mathrm{A}}{M_\mathrm{n}}\,,$$

or, with $M_\mathrm{n} = A_\mathrm{r} \cdot$ kg/kmol,

$$\frac{K}{\varrho} = 0.04 \cdot \frac{Z}{A_\mathrm{r}} \cdot \frac{\mathrm{m}^2}{\mathrm{kg}}\,. \tag{26.20}$$

Experimentally, however, the measured value (Sect. 26.6) was found to be:

$$\frac{K}{\varrho} = 0.02\,\frac{\mathrm{m}^2}{\mathrm{kg}}\,. \tag{26.15}$$

The comparison of (26.20) and (26.15) gives

$$Z = 0.5\,A_\mathrm{r}\,. \tag{26.21}$$

In words, this means that *in the inner shells of an atom with a moderate molar mass, the effective number Z of electrons is equal to $A_r/2$*. (A_r is the quantity *relative atomic mass*, a pure number proportional to the atomic mass of each element; formerly, it was called the *atomic weight*).[C26.8] This fundamental piece of knowledge of atomic structure is due to J.J. THOMSON (1906).

C26.8. The number A_r corresponds roughly to the nuclear mass number A, the total number of nucleons (protons and neutrons) in the nucleus of an atom of a given element (i.e. a given isotope). The statement made here thus means that the number of scattering electrons is half as large at the number of nucleons in the nucleus of the corresponding atom. In a neutral atom, the number of electrons equals the number of protons in the nucleus (Z), so this means that in lighter elements, the number of neutrons equals the number of protons in the nucleus. This is well known from nuclear physics.

26.8 Scattering as a Means of Producing and Detecting Polarized X-rays

In the visible spectral range, and in the neighboring regions, we can make use of RAYLEIGH scattering not only for the detection of linear polarization (Sect. 24.2), but also for producing it.

The polarization of radiation by means of scattering becomes fundamentally important only in the X-ray region. There, the other methods which can be used for ultraviolet, visible and infrared radiation, such as polarization prisms and foils, or reflection polarizers, can no longer be used. *In the X-ray range, scattering is the only*

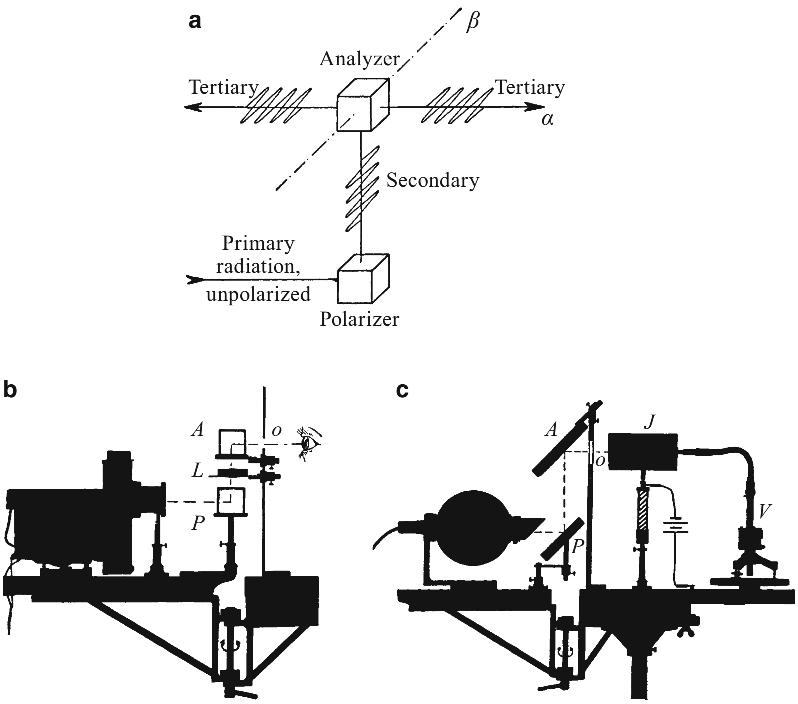


Figure 26.8 The production and detection of linearly-polarized light by means of scattering. a: Schematic, no tertiary radiation in the direction β. b: A demonstration experiment with visible light. c: A demonstration with X-rays (analyzer A fixed, polarizer P and source can be pivoted together on an arm around the vertical axis. The X-ray source requires AC at 220 V. J is an ionization chamber (Fig. 15.7), V is a static voltmeter with power supply and a light-beam pointer; L is a lens). A and P are cloudy water for visible light (compare Fig. 24.4), and for X-rays, they are made of materials with a small molar mass, for example paraffine. The flat shape is intended to reduce absorption losses. The openings o, which are not visible in the silhouette, are indicated by outlines; likewise, an insulating column is shown by shading. Of course, the eye of the observer at left could be replaced by a suitable radiometer.

method of polarization. However, this holds only in the characteristic wavelength range, as we have seen in Sect. 26.6. The scatterers which are used for producing and detecting the polarization of X-rays must contain only atoms of light or moderate molar mass. Figure 26.8 shows the procedure, both for the visible region and for X-rays.

The polarization of X-rays was detected for the first time in 1905, thus 10 years after RÖNTGEN's discovery of X-radiation. It was the first new characteristic of X-rays, one which was not discovered by RÖNTGEN himself and not contained in his original publications.

26.9 The Scattering of Visible Light by Large, Weakly-Absorbing Particles

Often, the scattering objects are not small compared to the wavelength of the light. Then the simple characteristics of RAYLEIGH scattering no longer apply. For example, the symmetry of the scattered radiation around the direction of the incident light is no longer present.$^{C26.9}$ Instead, we see mainly "forward scattering", i.e. scattering along the direction of the incident beam of light. For a demonstration, small sulfur particles in water are suitable. We use the setup shown in Fig. 26.9. The glass tube contains a solution of $Na_2S_2O_3$, with a small amount of H_2SO_4 added. This causes sulfur to precipitate in the form of small, solid suspended particles. The size of the particles increases in the course of several minutes. During this process, forward scattering becomes more and more prominent (Fig. 26.10).

C26.9. In order to see this symmetry in RAYLEIGH-scattered radiation, imagine that Fig. 26.4 were extended to three dimensions, as suggested in the last sentence in the figure caption.

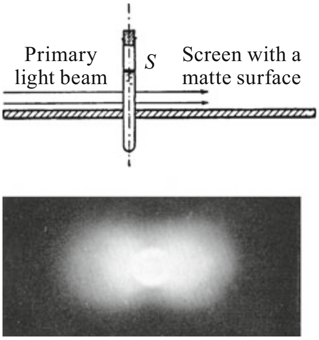


Figure 26.9 *Top*: The experimental setup for demonstrating scattering (about 1/6 actual size). The primary light beam passes above the matte-surface screen without touching it. The glass tube *S* contains suspended particles of sulfur in water. The screen is illuminated only by the scattered light. *Bottom*: The rough symmetry of the scattered radiation from *small* suspended particles. The primary light beam (red-filter light) is linearly polarized. Its plane of oscillation is parallel to the surface of the screen (photographic positive; cf. Fig. 26.4).

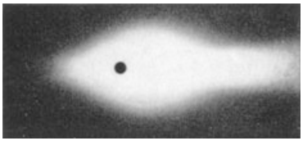


Figure 26.10 The lack of symmetry of the scattered radiation from *large* sulfur particles: "forward scattering", along the direction of the primary light beam, predominates (the setup is shown in Fig. 26.9, top; the light used is unpolarized, from an incandescent lamp. This image is about 1/10 actual size).

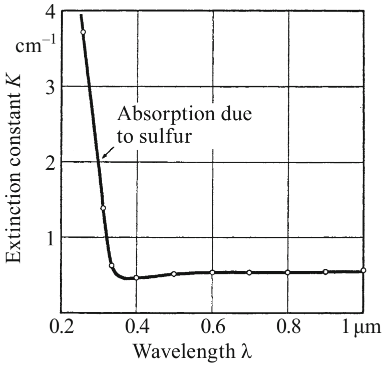


Figure 26.11 The influence of the wavelength on the extinction constant of the suspension of fine sulfur particles used as scatterer in Fig. 26.9. Below $\lambda = 350\,\text{nm}$, the sulfur begins to absorb the light strongly, i.e. the light is no longer scattered, but instead is converted to heat (see Sect. 27.13)

An additional important point is the dependence of the scattering on the wavelength. For large particles, RAYLEIGH*'s law* no longer holds, that is the wavelength dependence of the extinction constant:

$$K = \text{const} \cdot \frac{1}{\lambda^4}\,. \qquad (26.11)$$

The exponent becomes smaller and smaller, the larger the scattering particles$^{\text{C26.6}}$. In the example shown in Fig. 26.11, K has become practically independent of λ. The eigenfrequency ν_0, which depends on the size of the scattering particles (the "antenna length"), is much smaller than the frequency ν of the incident light.

From forward scattering, we continue on to diffraction, when the size of the particles on which the light is incident reaches the same order of magnitude as the wavelength of the light. This case can be readily demonstrated in a model experiment using water waves. The particles are constructed from individual 'building blocks', which are small steel balls of about 3 mm diameter, placed below the surface of the water. Each of these invisible "obstacles", when struck by the primary wave, becomes the source of secondary scattered wavelets. The wavelets interfere with each other, and this produces a *diffraction pattern* from the round particles. Figure 26.12 shows some snapshot images on the background of the primary waves. When the particles move or rotate, the well-defined preferential scattering directions disappear, and the superposition of different diffraction patterns with different shapes and orientations gives a washed-out diffraction pattern, concentrated along the direction of the incident primary waves.

We could consider the arrangements of the steel balls in Fig. 26.12 a and b as models for ring- and rod-shaped *molecules*; the balls themselves would be the atoms, and the waves would be X-rays. The directional distribution of the scattered waves, which interfere and are combined into a diffraction pattern, permits conclusions to be drawn about the structure of these 'molecules'.$^{\text{C26.10}}$

C26.10. An example of such an investigation is the discovery of the double-helix structure of DNA (F.H.C. Crick, J.D. Watson, and M.H.F. Wilkins, in *Nobel Lectures in Molecular Biology*, 1962 (Elsevier, NY 1977), pp. 147–215).

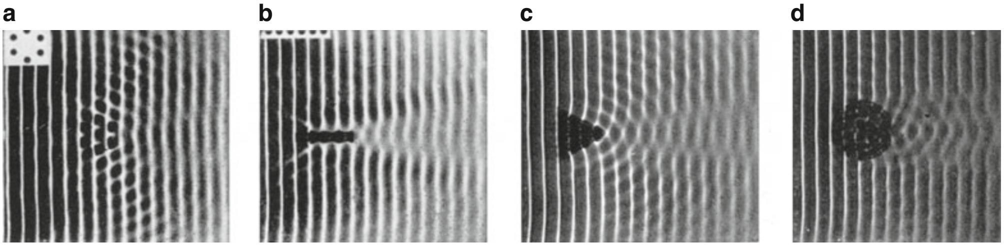


Figure 26.12 A model experiment showing the transition from scattering to diffraction by weakly-absorbing particles whose diameter is larger than the wavelength. In Parts a and b, the arrangement of the individual 'building blocks' (steel balls under the water surface) is shown at the upper left on the same scale as the main image. In Part c, the balls form a triangular object, and in Part d, it is circular

26.10 Diffuse Reflection from Matte Surfaces

What we have learned thus far about scattering will allow us to understand *diffuse reflection* from matte surfaces. Matte surfaces consist of fine, usually crystalline 'dust' particles or fibers (paper!), made of weakly-absorbing materials. Figure 26.13 shows an example.

We can distinguish three contributions to diffuse reflection:

First, the *reflection* by numerous extremely small and disoriented 'mirror surfaces', which are the boundary surfaces of the crystallites. The radiant intensity of the light reflected from these micro-mirrors obeys LAMBERT's cosine law up to moderate values of the angle of incidence (Sect. 19.2). Only at large angles of incidence do the directions pointing away from the light source become predominant: In these directions, the rays from mirrors that were struck at a very flat angle are concentrated, and, according to FRESNEL's formulas (Sect. 25.8), they are more intense than the rays from mirrors struck at more nearly perpendicular incidence.

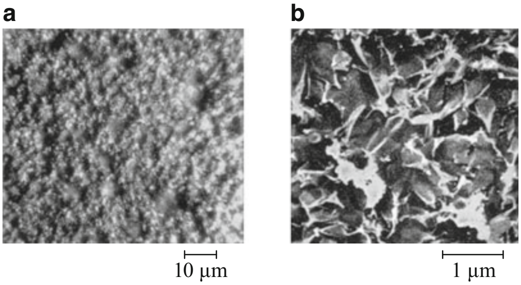


Figure 26.13 Photomicrographs of a matte zinc oxide layer, prepared by condensation from the vapor (at left, an optical image using visible light; at the right, an electron microscope image)

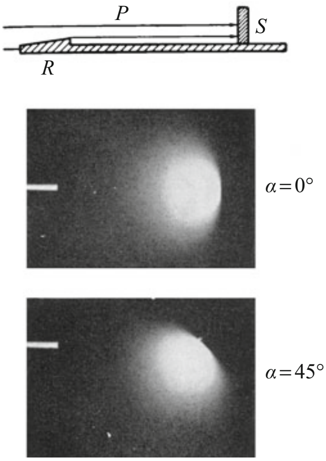


Figure 26.14 Demonstrating LAMBERT's cosine law for scattered radiation from the matte surface of a piece of chalk

To demonstrate LAMBERT's cosine law in the case of diffuse reflection, we can use the arrangement shown in Fig. 26.14. The primary radiation P is in the form of a collimated beam. It grazes a flat ramp R (at the end of a board, shown as a shaded cross-section) and thereby indicates its direction and diameter. Then it falls on the matte, planar surface of a piece of chalk (S). The radiation scattered there by diffuse reflection illuminates the board and produces a second diffuse reflection. Its rays can be seen directly or photographed by a camera. Both views are perpendicular to the board. Chalk exhibits a nearly "ideally-diffuse" reflection: The radiation is itself still symmetric around the surface normal of the chalk, even when the angle of incidence α of the primary beam is $\approx 45°$. Paper and porcelain also produce very diffuse reflections. Their glazing has no effect on this; it only produces additional specular reflections which remain within the plane of incidence.

The *second* contribution to diffuse reflection is a genuine type of *scattering*, the secondary radiation coming from tiny powder crystallites. For larger scattering particles, it is confined mainly to the direction of the incident light and a narrow cone which surrounds it. This forward scattering is generally directed into the powder layer and produces multiple scattering by the deeper layers within its interior. This again obeys LAMBERT's cosine law for the rays which finally emerge from the surface. Only when the angle of incidence α becomes large, that is for grazing incidence, is the direction away from the light source again preferred (Fig. 26.15).

A *third* contribution to diffuse reflection comes about because even *matte* surfaces act as *good mirrors* at large angles of incidence. We offer two examples of this:

Figure 26.15 was obtained using the zinc oxide layer already shown in Fig. 26.13. It shows the distribution (as a polar diagram) of the secondary radiation for an angle of incidence of $\alpha = 80°$. We can see strong forward scattering, superposed onto a specular reflection (giving the sharp peak Sp). The radiant intensity of the light reflected from the matte surface exceeds

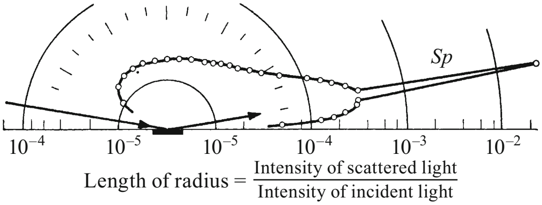


Figure 26.15 The forward scattering from a matte zinc oxide surface, superposed onto a specular reflection *Sp* (H.U. Harten, *Zeitschrift für Physik* **126**, 27 (1949))

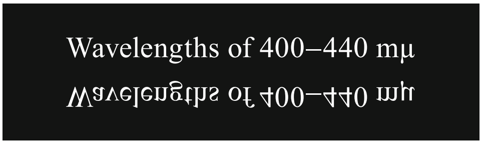


Figure 26.16 *Below*, a direct photograph of printed type; *above*, a reflected image, after reflection at a grazing angle from a matte glass plate (angle of incidence $\alpha = 89.5°$). Instead of the printed block letters, we could also image a slit onto a screen with a lens and use a matte glass plate as mirror, at grazing incidence. With increasing angle of incidence, the screen at first becomes brighter due to forward scattering. On this bright background, we see the reflected image of the slit, at first weak and reddish, then becoming brighter and whiter

that of the scattered light by a factor of around a hundred (logarithmic scale!).

Figure 26.16 shows two images of the same printed block letters. The lower one was photographed directly, while the upper was reflected at grazing incidence ($\alpha = 89.5°$) from a matte glass plate.

The explanation for this mirror-imaging is not hard to find: The uppermost ‘peaks’ of the crystallites on the matte surface act like a two-dimensional point grating with statistically-distributed grating constants. The zeroth order has the same direction for all of these ‘partial gratings’, i.e. the direction corresponding to specular reflection. The flatter the incident light beam relative to the surface (grazing incidence), the smaller the grating constants become, due to perspective foreshortening. This eliminates the higher orders, and finally the whole illumination comes from the zeroth-order radiant intensity diffracted by the grating points.

Exercises

26.1 In our quantitative treatment of scattering, we mentioned harmonic oscillators (in Sect. 26.2). Consider a spring pendulum with a mass m (deflection l, logarithmic decrement Λ). It is excited to forced vibrations by a periodic force $F_\mathrm{p} = F_0 \cos(2\pi\nu t)$ (see Eq. (26.2)). Their amplitude l_0 and their phase shift φ are given by Eqns. (26.1) and (26.3). These two equations are to be derived from NEWTON's fundamental equation, $F = m\mathrm{d}^2 l/\mathrm{d}t^2$ (equation of motion, Vol. 1, Chap. 4).

a) In a first step, show that the equation of motion for a freely-vibrating spring pendulum which is damped by a frictional force $F_\mathrm{R} = -\alpha\,\mathrm{d}l/\mathrm{d}t$ that is proportional to its velocity, is solved by the equation $l = l_0 e^{-\delta t}$, with $\delta = (1/2)(\alpha/m)$. Find the relation between α and Λ.

b) Write the equation of motion for forced vibrations, and solve it using a complex trial solution, $l = l_0 e^{i(2\pi\nu t)}$. Derive from this Eqns. (26.1) and (26.3). (See the footnote on complex numbers in Sect. 25.10).

(For Sect. 26.2)

Dispersion and Absorption 27

27.1 Preliminary Remark and Overview of the Chapter[C27.1]

The refractive index n depends on the wavelength λ of the radiation; it exhibits "dispersion". Dispersion is closely related to the absorption of radiation, and it, in turn, depends strongly on the wavelength. In the following Sects. 27.2 through 27.5, we will consider the empirical evidence relating dispersion and absorption. Then we will treat quantitatively the wavelength dependence of refraction and absorption. This will show close ties to our quantitative treatment of scattering in Chap. 26.

[C27.1] This chapter goes well beyond the topic of optics and deals with many particular details from solid-state physics. These are intended to illustrate general applications of the optical concepts *dispersion* and *absorption*, which is the goal of this chapter.

27.2 The Wavelength Dependence of Refraction and Extinction

Let us recall Sect. 25.2: We refer to the *extinction* constant K and the *extinction* coefficient k as the *absorption* constant and absorption coefficient when the effects of scattering on the measured extinction are negligible. The basic facts can be most clearly illustrated by a graphical representation. To represent the refractive index, we draw "dispersion curves". For the extinction, depending on the application, we can represent the same measurements in two ways: Either in terms of the extinction constant K, or, in the case of strong absorption, in terms of the extinction coefficient k. The latter compares the average penetration depth of the radiation (i.e. $w = 1/K$) with its wavelength, as we have seen in Eq. (25.3). Thus, the extinction coefficient k naturally has a very different form from the extinction constant K across the spectrum.

Unfortunately, we have only fragmentary knowledge of both the dispersion curves and the extinction curves for most substances. This lack of knowledge is however minimal for the simplest of solids, the regular crystals of the alkali halides. Therefore, we begin in Fig. 27.1 with measurements on NaCl (rock salt). We first focus our attention on its refractive index. In the X-ray region, i.e. $\lambda <$ ca. $5 \cdot 10^{-8}$ m, the refractive indices are all slightly less than 1 (Sect. 27.9). The tiny deviations from 1 cannot be seen on the ordinate scales of the figures. In the region of longer wavelengths, the refractive index approaches

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_27

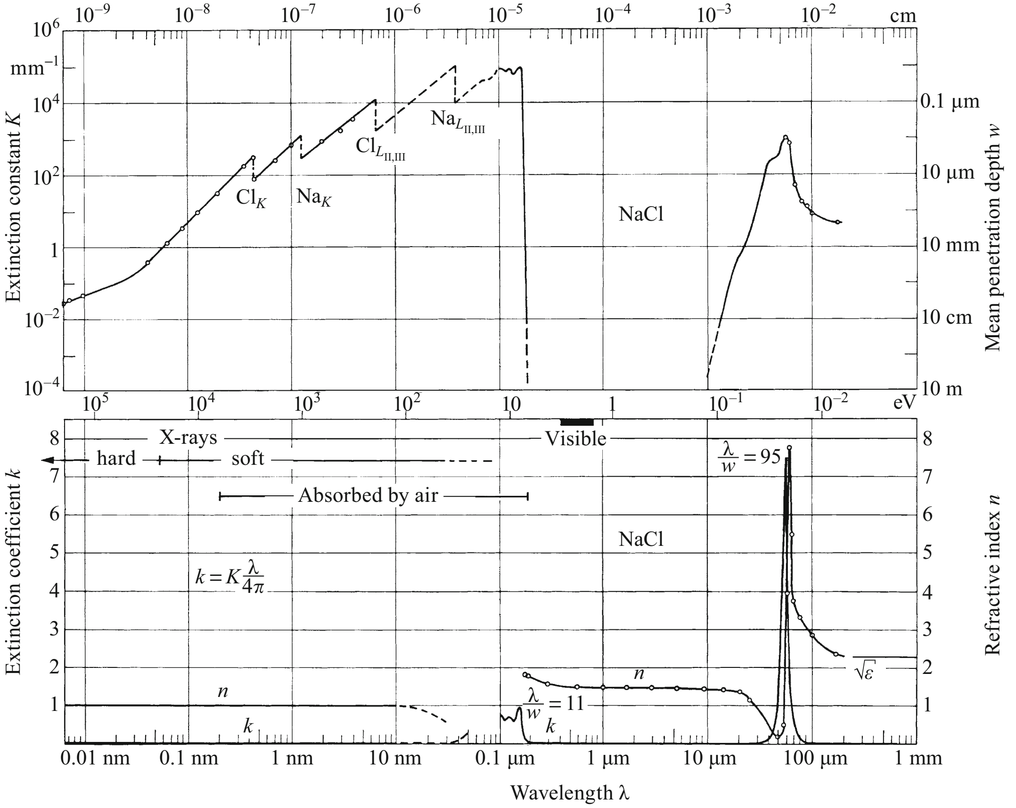


Figure 27.1 Refraction and extinction (absorption) of light by an NaCl crystal between $\lambda = 6 \cdot 10^{-12}$ m and 1 mm, thus over a range of about 28 octaves.$^{\text{C27.2}}$ The extinction coefficient k reaches significant values only in two narrow wavelength regions, namely from about 0.04 to 0.2 μm and from about 30 to 90 μm. In these regions, the highest values of the ratio λ/w are recorded. The smallest penetration depth that occurs, $w \approx$ 0.01 μm, is around 30 times larger than the spacing of the lattice planes in the crystal. The occurrence of the absorption "edges" Cl_K etc. in the X-ray region is due to the fact that with increasing wavelength, the energy of the light quanta is at some point no longer sufficient to knock the electrons out of their inner shells

C27.2. The abscissa shows the wavelength, above in cm and below in (nano, micro, milli)-meter. On the center abscissa axis, the associated quantum energy is quoted in electron volt (eV), corresponding to $E = h\nu = hc/\lambda$ (h is PLANCK's constant $= 6.626 \cdot 10^{-34}\,\text{W s}^2$, c (velocity of light) $= 2.998 \cdot 10^8$ m/s, 1 eV $= 1.602 \cdot 10^{-19}$ W s).

the square root of the statically-measured dielectric constant ε, that is $n = \sqrt{\varepsilon}$ (see Sect. 12.8). In most spectral regions, the refractive index n increases with decreasing wavelength; the dispersion is then termed 'normal'. In some spectral regions, however, n *decreases* with decreasing wavelength. Then the dispersion is called 'anomalous', i.e. it deviates from the rule.

The distinguished regions in the dispersion curves, that is the regions where n is varying most rapidly, and those where it has an 'anomalous' wavelength dependence, coincide with those regions where the absorption coefficient k is largest. This is documented in Fig. 27.2 with five additional examples. At the boundary of an absorption band, the rate of change of the refractive index with wavelength,

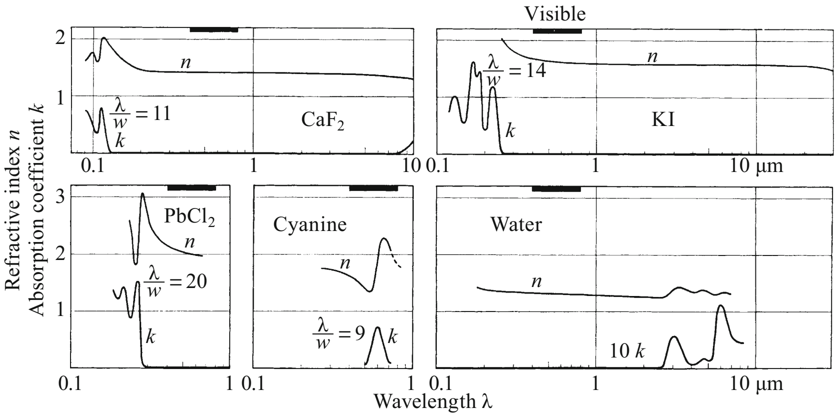


Figure 27.2 Five additional examples of dispersion and absorption. The birefringence of PbCl$_2$ cannot be seen on the scale of the figure. The values of k for water are magnified by a factor of 10, as indicated

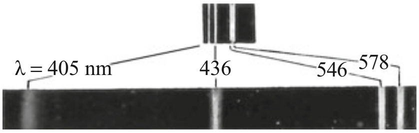


Figure 27.3 A demonstration experiment illustrating the strong dispersion in the spectral region just before the steep onset of self-absorption: The visible part of the Hg line spectrum, obtained under similar conditions with two different 60° prisms; below, a prism made from a single crystal of ZnO (embedded in water), and above, a quartz prism (E. Mollwo, *Zeitschrift für Angewandte Physik* **6**, 257 (1954))

i.e. its *dispersion* dn/dλ, can become very large. This is shown in Fig. 27.3 with the aid of a prism made of ZnO.

The relationship between dispersion and absorption can be demonstrated by an impressive experiment. For this purpose, neither solids nor liquids are suitable[1]; we must use vapors, gases or very dilute solutions. Sodium vapor is most convenient. Figure 27.4 shows a suitable experimental arrangement. It uses a prism P to project the continuous spectrum of a sodium-arc lamp onto a horizontal screen. An iron tube R filled with Na vapor is placed directly behind the projection lens L_1. It has glass windows at each end and is evacuated. Then the sodium is evaporated in the middle of the tube. A low pressure of H$_2$ gas and air cooling prevent the windows from being clouded by condensing Na vapor. The vapor produces a strong extinction around $\lambda = 589\,\text{nm}$. The horizontal spectrum is interrupted

[1] The justification of this statement will be seen later in Eq. (27.7). n attains high values only when the difference of the *squared* frequencies, that is $\nu_0^2 - \nu^2$, is very small. In solids and liquids, with their *broad* absorption bands, this leads into regions where the material is opaque.

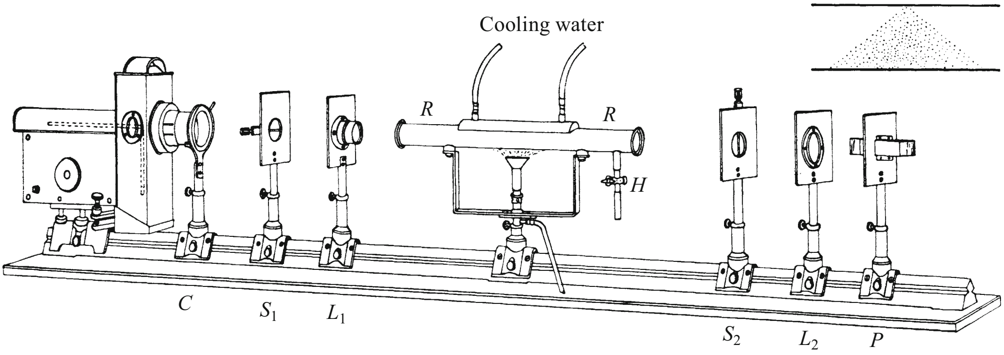


Figure 27.4 A demonstration of anomalous dispersion in Na vapor (A. KUNDT, 1880, improved by R.W. WOOD, 1904.[C27.3] S_1 is a horizontal slit, S_2 a vertical slit, and P is a direct-vision prism[C27.4]). The vapor prism (schematic in the inset at upper right) deflects waves with a refractive index of $n > 1$ downwards, while waves with a refractive index of $n < 1$ are deflected upwards. In the example shown in Fig. 27.5 (*lower image*), the lower end of the slit S_2 is imaged upwards onto the screen. A cylindrical lens between R and S_2 improves the visibility

C27.3. See R.W. Wood, *Physical Optics*, McMillan Publishers, NY, 3rd ed. (1934), p. 492.

C27.4. In a "direct-vision" prism, a combination of strongly and weakly refracting glasses is employed, so that at a certain wavelength, a light beam is not deflected on passing through the prism.

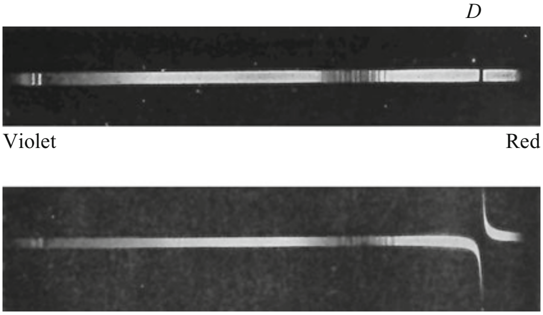


Figure 27.5 The anomalous dispersion of Na vapor, demonstrated with the apparatus in Fig. 27.4 (photographic positives). The absorption bands which are visible in addition to the D band are due to Na molecules. Because of their low particle-number densities, they have no noticeable influence on the refractive index of the vapor.

by an extinction band D (Fig. 27.5, *upper image*). Here, the contribution due to absorption is larger than that due to scattering. As a result, we nearly always speak of *absorption* bands or lines.

Following this preliminary experiment, not only the ends but also the upper part of the tube are cooled. This causes the cloud of Na vapor in the tube to take on a prismatic shape (schematic in the inset at the upper right in Fig. 27.4). At the hottest point, i.e. at the bottom of the center of the tube, the density of the vapor is high; upwards and towards both ends, it decreases. This 'vapor prism' leaves most of the spectrum unchanged. In these unchanged spectral regions, the refractive index of the Na vapor is practically equal to 1. On both sides of the absorption band, however, the light is deflected vertically. On the red side, the deflection is downwards on the slit S_2, that is the refractive index is > 1. On the violet side of the band, the deflection on the slit S_2 is upwards, that is the refractive index is < 1. The spectrum thus forms a colored curve consisting of two branches (Fig. 27.5, *lower image*).[C27.5] The shape of this curve corresponds to the dispersion curve of Na vapor on both sides of the extinction (absorption) band. The section of the curve *within* the band is missing. It can be seen only with moderate absorption and then only by direct, subjective observation.

[C27.5] A colored image can be found in the book by R.W. Wood; see Comment C27.3.

27.3 The Special Status of the Metals

We return to the important Fig. 27.1. The smallest values of the extinction constant K, i.e. the greatest penetration depths w, can be found in the visible and the neighboring spectral regions, in particular in the infrared. In these regions, the mean penetration depth can be up to many meters and greatly exceeds those for all other radiations, in particular for X-rays. An exception to this rule is provided by the metals. This is shown in Fig. 27.6 for silver. The figure spans a wavelength range of 16 orders of magnitude.

The extinction constant K has very high values over the entire infrared and visible spectral regions; the extinction processes acting there extend into the ultraviolet. Details will be given in Sect. 27.17.

Values for n and k are shown for two important metals in Fig. 27.7. The absorption coefficients k increase to high values from the ultra-

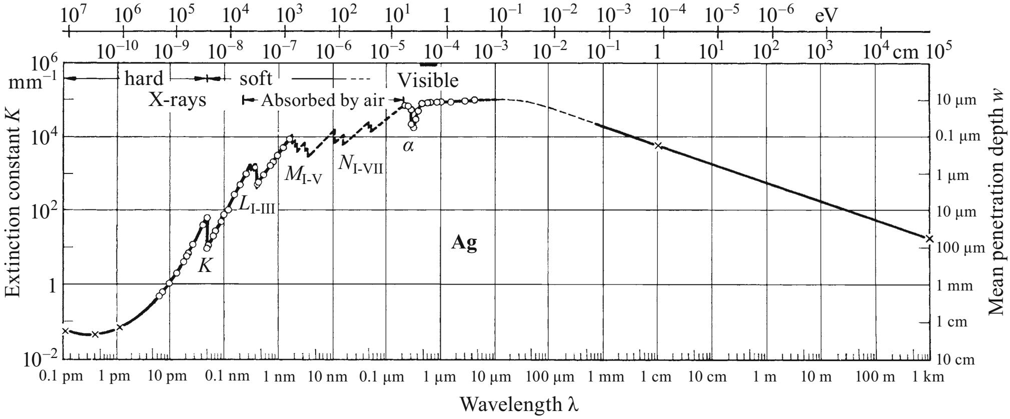


Figure 27.6 The extinction (absorption) spectrum of a metal (silver) between $\lambda = 10^{-13}$ m (0.1 pm) and $\lambda = 1$ km (the scale of the abscissa is half that of Fig. 27.1). The gap in the extinction curve for NaCl as seen in Fig. 27.1 between 0.2 μm and 20 μm is not present in this curve. The small minimum α at $\lambda = 0.32$ μm is by no means comparable to that gap. The mean penetration depth w attains a value of only 50 nm here. The points marked with 'x' are calculated. For Al, the minimum of the extinction constant is at $\lambda = 6 \cdot 10^{-14}$ m. There, the penetration depth $1/K = w = 17$ cm. (Upper scale: 1 eV $= 1.602 \cdot 10^{-19}$ W s)

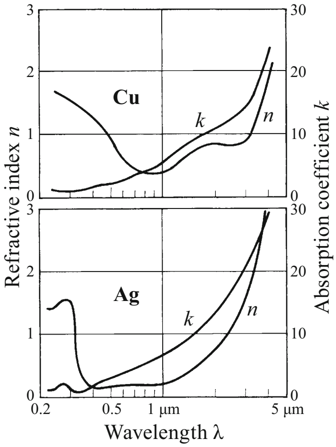


Figure 27.7 The optical constants n and k for silver and copper. The scatter among the values is still rather large, even in the best measurement series available today. Further examples are given in Fig. 27.22.

C27.6. The phase velocity of light within materials can indeed become higher than the vacuum velocity of light, c. This does not mean that information can be transferred at a velocity faster than c. For information transfer, the *group velocity* is the decisive quantity (see Sects. 23.3 and 12.11), and as we have seen, it is always less than c.

violet towards longer wavelengths. At $\lambda = 4\,\mu\text{m}$, for example for silver, we find $k \approx 30$. The mean penetration depth w here is thus equal to $\frac{1}{400}\lambda$. The smallness of the refractive index n is also often notable in metals. For silver, it decreases to 0.16. Thus, the phase velocity increases up to nearly $20 \cdot 10^8$ m/s instead of only $3 \cdot 10^8$ m/s for light in vacuum.C27.6

27.4 'Metallic' Reflectivity

For the reflectivity R at perpendicular incidence, we can apply BEER's formula:

$$R = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2}\,. \qquad (25.37)$$

If the summand k^2 is predominant in the numerator and the denominator, then we observe the high reflectivity ($R \approx 1$) which is characteristic of materials with metallic bonding in the visible spectral region (Sect. 25.11). Figure 27.8 shows some examples which are of practical importance (more on this in Sect. 27.17). Metallic bonding is however by no means the only reason for very large values of the absorption coefficient k. Values of k on the order of 1 in the ultraviolet can be found for the majority of solid and liquid substances. Some examples are given in Fig. 27.2. In the case of dyes, e.g. cyanine or brilliant green (see Table 25.1), and for some semiconductors, the absorption coefficient k attains large values already in the visible. As a result, some semiconductors such as germanium, silicon, stibnite (Sb_2S_3) etc. are indistinguishable from metals to the unaided eye. However, the semiconductors lack the large absorption coefficients in the infrared which are characteristic of substances with metallic

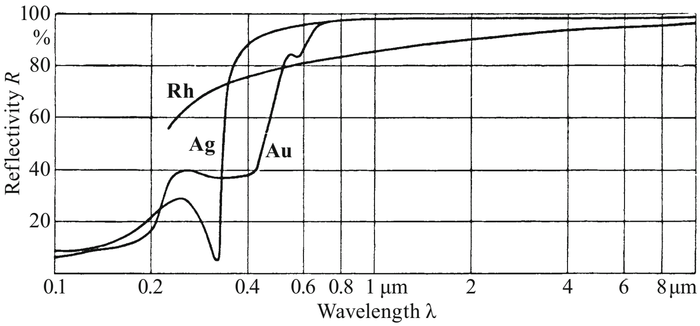


Figure 27.8 The wavelength dependence of the reflectivities of gold, silver, and rhodium. Rhodium is particularly suitable for mirror surfaces without glass protection, due to its chemical inertness. In addition, thin, transparent rhodium mirrors attenuate all the wavelength regions of the visible spectrum (400 to 700 nm) to practically the same degree (they act as "grey filters"). In the minimum at $\lambda = 316$ nm, $R = 4.2\,\%$ for silver. The values for the alkali metals are still smaller: at $\lambda = 254$ nm, $R = 2.6\,\%$ for potassium and $\approx 1\,\%$ for rubidium and cesium.

bonding, and which are caused by their special type of electronic conduction (Sect. 27.17).

In the case of germanium, for example, k is already vanishingly small at $\lambda = 3\,\mu$m. Thus, blocks of germanium a few cm thick look like pieces of metal. Nevertheless, they are transparent to infrared radiation, apart from the considerable reflection losses due to their large refractive index of $n = 4$.

This can be demonstrated by a very surprising experiment. It makes use of the setup shown in Fig. 25.1, and shows quite clearly that *whether metallic bonding is present or not can never be determined by the unaided eye; only absorption measurements in the infrared are conclusive*.

Finally, crystals with typical ionic bonding, such as the alkali-metal halides, exhibit extreme values of n and k in the infrared (compare Fig. 27.1, bottom). As a result, such crystals have a rather large reflectivity there. Figure 27.9 shows four examples. The wavelength scale is three times larger than in Fig. 27.1. These reflectivity maxima are called *residual rays*. Their position is determined by both n and by k. Therefore, their maxima are only approximately at the same positions as those in a plot of the absorption coefficient k.

The unusual name 'residual rays' is related to the method of their first observation. HEINRICH RUBENS (1865–1922) passed the radiation from an incandescent light heated by a gas burner back and forth several times with reflections between two crystal plates and then detected it using a radiometer (radiation thermopile). The remaining ("residual") rays consisted of practically only those waves from the spectral regions of the reflection maxima. These "residual rays" are absorbed by thin mica or glass plates,

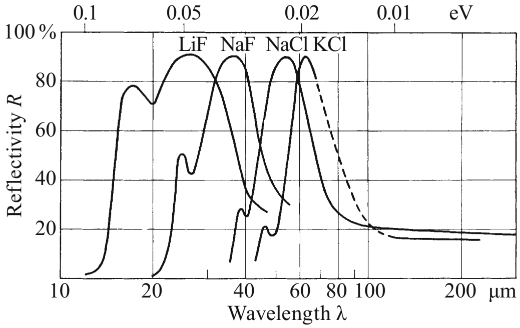


Figure 27.9 Residual rays exhibited by four alkali-metal halide crystals (the bands are not simple Gaussian curves, in contradiction to earlier reports)

but can pass through thick layers of paraffine, etc. (This is readily shown as a demonstration experiment, most simply using thin slabs of LiF or CaF_2).

27.5 The Penetration Depth of X-rays

The penetration depths of X-rays are greater than those of visible light only in metals (Fig. 27.6). In all other materials (for example NaCl; cf. Fig. 27.1), X-rays do not have anything like the enormous penetration depths which are observed with light from the visible and the neighboring infrared spectral regions.

The importance of X-rays for medical and technical applications is however not based on their great penetration depths, but rather on something quite different: *The refractive index for X-rays is very close to 1* (Sect. 27.9). *As a result*, X-rays *experience no diffuse reflection* in cloudy, inhomogeneous materials such as muscle tissue, bones, wood etc. They are unaffected by the innumerable boundaries between the individual components of inhomogeneous materials. Visible light, in contrast, with refractive indices of about 1.5, is exceedingly sensitive to internal boundary surfaces. The 'head' on light Pilsner beer is quite opaque to visible light, but completely transparent to X-rays.

"The 'head' on light Pilsner beer is quite opaque to visible light, but completely transparent to X-rays".

The lack of diffuse reflection in the X-ray range does not imply that there is no scattering there (see Sects. 26.6–26.8). Scattering plays a significant role even for hard X-rays ($\lambda < 10^{-11}$ m). It is caused in this spectral region by the COMPTON effect (see for example the 13th edition of "*Optik und Atomphysik*", Chap. 17. English: See e.g. hyperphysics.phy-astr.gsu/hbase/quantum/comptint.html or https://en.m.wikipedia.org/wiki/Compton_scattering), and at still shorter wavelengths, by nuclear processes as well.

27.6 The Explanation of Refraction in Terms of Scattering

From reading Sects. 27.2 through 27.5, we are now familiar with the more important empirical facts about refraction and extinction. Here, we want to interpret them and describe them quantitatively. In Sects. 27.6 through 27.11, we treat refraction; and in Sects. 27.12 through 27.21, we deal with that part of the extinction which is due to absorption.

We return to Fig. 26.12 d. There, the model scattering object is transparent. We can follow the course of the waves within it, although with some difficulty. We observe a picture like the one sketched in Fig. 27.10: The waves propagate more slowly in the region where secondary-wave sources are present than outside the object; the wave crests are noticeably delayed within the object. Or, expressed differently, the circular region has acquired a *refractive index* unequal to 1 due to the presence of secondary-wave sources in its interior. This basic fact will be verified by an impressive demonstration experiment.

The best-known effects of refraction are exhibited by *lenses*. Thus, in Fig. 27.11 a, we show a group of "secondary-wave sources" arranged in a lens-shaped pattern. The scattering "atoms" are again small steel balls below the surface of the water in a wave trough. They are disordered, and their diameters and spacings are again smaller than the wavelength. In Fig. 27.11 b, water waves with straight wavefronts pass at a small angle of inclination through a wide slit. The slit cuts out a collimated (parallel-bounded) beam of waves; diffraction can be clearly seen.

In Fig. 27.11c, the "lens" has been placed in the opening of the slit. The result is that the previously collimated wave beam is now focussed onto an image point. Now, there can be no further doubt: The waves pass through the region containing secondary sources with a *reduced phase velocity*. The region with secondary sources has a refractive index n. We compute it using the elementary lens formula

$$(n-1)\frac{2}{R}=\frac{1}{f'} \qquad (16.12)$$

(R is the radius of the lens surface; in Fig. 27.11a, $R = 7\,\text{cm}$ and $f' \approx 8.5\,\text{cm}$)

and obtain $n = 1.4$.

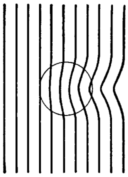


Figure 27.10 The occurrence of a phase shift due to secondary waves (sketched from Fig. 26.12d)

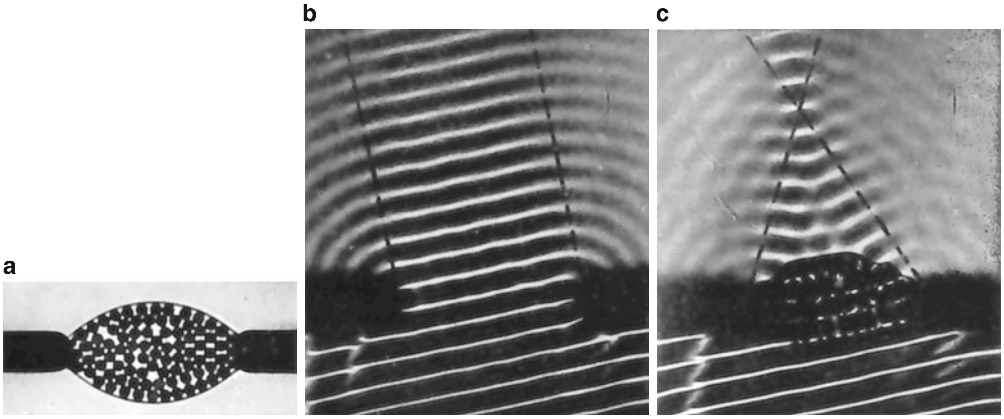


Figure 27.11 Water waves exhibit the occurrence of *refraction* through phase-shifted secondary waves

The explanation is readily found. The waves which propagate within and behind the lens are the resultants of all the secondary waves which are produced by scattering, together with the primary waves. The primary waves generate secondary waves, and these generate tertiary waves, etc. The resultant waves have an overall phase velocity which is less than that of the individual wave components. Therefore, each individual secondary wave produced by scattering must have a negative phase shift δ' relative to the primary waves. *The phase shift δ' of the secondary waves generated by scattering causes the refraction*.

27.7 The Qualitative Interpretation of Dispersion

The wavelength dependence of the refractive index n exhibits a very characteristic form in the neighborhood of certain distinguished wavelengths or frequencies (Figs. 27.1 and 27.2). We repeat it schematically in Fig. 27.12. This dependence of the refractive index on the wavelength or the frequency is readily interpreted in a qualitative way. To this end, we turn again to our model experiments with mechanical waves.

In Fig. 27.11, the secondary-wave sources consisted of small *rigid* balls beneath the surface of the water. Now, imagine that these secondary-wave sources are replaced by objects capable of *vibrations*, i.e. *resonators*; for example, by "breathing spheres" (Vol. 1, Sect. 12.25). Their eigenfrequency is denoted by ν_0. The incident primary waves have a frequency of ν and excite the resonators to forced oscillations. Then both the forced amplitudes l_0 and also the phase differences between the resonators and the primary waves are

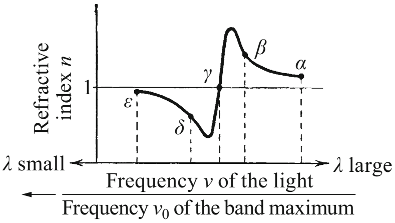


Figure 27.12 Schematic of a dispersion curve in the neighborhood of an optical eigenfrequency (band maximum; the points α to ε correspond to the images A–E in Fig. 27.13; see the text)

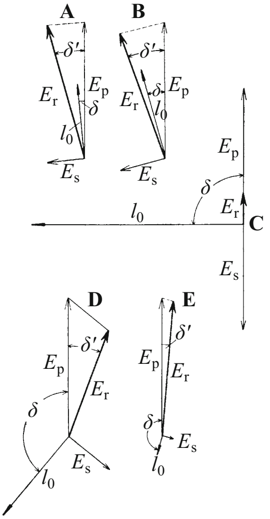


Figure 27.13 The occurrence of dispersion as a result of phase-shifted secondary waves (the time increases in the clockwise sense)

determined by the ratio ν/ν_0 (Sect. 26.2, see also Vol. 1, Fig. 11.42). *Furthermore, the amplitude of each secondary wave is shifted relative to the amplitude l_0 of the secondary-wave source by $-90°$*[2].

We thus arrive at the simple vector diagrams shown in Fig. 27.13 A–E. The notation there is:

E_p is the amplitude of the primary wave;

l_0 is the amplitude of the forced oscillations; their relative values can be read off from Fig. 11.42 in Vol. 1 (e.g. for $\Lambda = 1$);

δ is the phase angle between l_0 and E_p; it can likewise be read off from Fig. 11.42 in Vol. 1 (with $\Lambda = 1$);

E_s is the amplitude of the secondary waves emitted by the resonators;

E_r is the resultant wave amplitude from the superposition of the primary and secondary waves; and

δ' is the phase angle between E_r and E_p. The time and the phase angles δ and δ' are taken as positive in the clockwise sense in the diagrams.

[2] This is a simplifying hypothesis. In fact, this phase difference of $-90°$ results from the summation of all the secondary waves along the path of the primary waves.

In Part A of Fig. 27.13, $\nu \ll \nu_0$ and δ is very small. δ' takes on a small *negative* value. This means that the resultant waves are slightly delayed relative to the primary waves, or the refractive index n is somewhat larger than 1. It is drawn in Fig. 27.12 as the point α.

In Part B, $\nu < \nu_0$, e.g. $\nu = \frac{1}{2}\nu_0$; δ has increased to around $-15°$. δ' remains *negative*, but its magnitude is larger. This means that the refractive index n is also larger: Point β in Fig. 27.12.

In Part C, $\nu = \nu_0$, that is $\delta = -90°$. The resultant amplitude E_r (the difference $E_p - E_s$) has the same direction as E_p. Then $\delta' = 0$ or $n = 1$ (Point γ).

In Part D, $\nu > \nu_0$, e.g. $1.25\ \nu_0$, and $\delta = -140°$. Then, δ' takes on a *positive* value. The resultant amplitude E_r leads the primary amplitude E_p. This means that the refractive index is smaller than 1 (Point δ).

Finally, in Part E, $\nu \gg \nu_0$, and δ is almost $-180°$. δ' has remained *positive*, but its magnitude has decreased; n is approaching the value 1, but is still smaller than 1 (Point ε).

Figure 27.12 illustrates a typical dispersion curve. It exhibits qualitatively the same features as the curves observed in optical experiments. The distinguished wavelength (point γ) corresponds in the optical measurements to the maximum of an extinction band.

27.8 A Quantitative Treatment of Dispersion

Quantitatively, the treatment in the previous section is rather unsatisfactory. In particular, it distinguishes only the exciting primary wave from the excited secondary waves. In reality, however, the secondary waves for their part excite tertiary waves and so forth. Only the entirety of all these waves finally leads to the resultant. The summation is computationally not simple, but it can be carried out. In general, however, one avoids that difficulty by using the following procedure:

We assume that per molecule[3], there is one oscillating, bound electron; its eigenfrequency is denoted by ν_0. It can undergo forced oscillations under the action of a periodic force of amplitude $F_0 = e \cdot E_0$. Its oscillation amplitude l_0 is found from Eq. (26.1) to be proportional to E_0, the amplitude of the primary waves, and inversely proportional to the mass m of the electron and furthermore dependent on the frequency ν of the primary waves. An oscillating dipole is thus produced, and its electric dipole moment has the amplitude

$$p_0 = e \cdot l_0 = E_0 \frac{e^2}{m} f(\nu) \tag{27.1}$$

[3] Here, as always, 'molecule' refers to the smallest independent unit; this could often be also atoms or ions.

(the frequency-dependent expression $f(\nu)$ follows from a comparison with Eq. (26.1)).

The quotient

$$\frac{p_0}{E_0} = \frac{e^2}{m} f(\nu) = \alpha \tag{27.2}$$

is the *electric polarizability* of the molecules (see Sect. 13.9) at the high frequency of the light waves.

In our discussion of RAYLEIGH scattering, it was assumed that $\nu \ll \nu_0$. This made the polarizability α *independent* of the exciting frequency (Sect. 26.5). Therefore, α could be computed from the statically measured (i.e. at $\nu = 0$) dielectric constant ε. In Sect. 26.5, we used the CLAUSIUS-MOSSOTTI equation for this purpose (Eq. (13.28)):

$$\frac{p}{E} = \alpha = \frac{3\varepsilon_0}{N_\mathrm{V}} \left(\frac{\varepsilon - 1}{\varepsilon + 2} \right) \tag{27.3}$$

(ε_0 is the electric field constant, and N_V is the number density of the polarizable molecules).

It takes into account the influence of the surroundings on the polarizability of the molecules.

Now, we follow the reverse path. We drop the limitation $\nu \ll \nu_0$ and thus make α *dependent* on ν (Eq. (27.2)); we insert the α values into Eq. (27.3) and then calculate for each value of the exciting frequency ν a particular value of ε. Thus we obtain a dielectric "constant" which *depends on the frequency* ν.

Finally, we come to the decisive step: According to MAXWELL, for long-wavelength electromagnetic waves (Sect. 12.8),

$$n = \sqrt{\varepsilon}\,; \tag{27.4}$$

here, ε refers to the statically-measured dielectric constant, i.e. at $\nu = 0$.

We now apply this same relation to light waves, but *at every frequency ν, we use the frequency-dependent dielectric constant* especially calculated for that frequency, in order to compute the refractive index n for light of the frequency ν. In this manner, the dependence of the refractive index n on ν or λ can be quite satisfactorily reproduced.

We will now carry out this plan quantitatively. We again take Eq. (27.1) for the induced dipole moments of the molecules, but actually compute l_0 using Eq. (26.1). We avoid the frequency range close to the eigenfrequencies ν_0; the regions $\nu < 0.7\ \nu_0$ and $\nu > 1.4\ \nu_0$ suffice. In these regions, the forced deflections l_0 are practically independent of Λ, the logarithmic decrement ($\Lambda \leq 1$). Thus, we can neglect the second summand in the denominator and obtain

$$l_0 = \frac{1}{4\pi^2} \cdot \frac{e}{m} E_0 \frac{1}{\nu_0^2 - \nu^2} \tag{27.5}$$

or

$$\alpha = \frac{e\,l_0}{E_0} = \frac{p_0}{E_0} = \frac{1}{4\pi^2} \frac{e^2}{m} \cdot \frac{1}{\nu_0^2 - \nu^2}\,. \tag{27.6}$$

This value of the *frequency-dependent polarizability* α is now inserted into Eq. (27.3). We write n^2 in place of the frequency-dependent ε and finally arrive at:

$$\frac{n^2 - 1}{n^2 + 2} = \frac{1}{12\pi^2 \varepsilon_0} \frac{e^2}{m} \cdot N_V \cdot \frac{1}{\nu_0^2 - \nu^2} = 26.9\,\frac{\mathrm{m}^3}{\mathrm{s}^2} \cdot N_V \frac{1}{\nu_0^2 - \nu^2} \tag{27.7}$$

(ε_0 is the electric field constant $= 8.86 \cdot 10^{-12}$ A s/V m, $e = 1.6 \cdot 10^{-19}$ A s, m is the mass of the electron $= 9.11 \cdot 10^{-31}$ kg, and N_V is the number density of the polarizable molecules).

Equation (27.7) assumes that there is only one eigenfrequency ν_0 and *one* electron per molecule. In reality, each molecule possesses a whole series of optical eigenfrequencies (numbered by the index i) and often also several effective electrons (number denoted by b). Thus, we must write a sum in place of Eq. (27.7), namely

$$\frac{n^2 - 1}{n^2 + 2} = 26.9\,\frac{\mathrm{m}^3}{\mathrm{s}^2} N_V \sum_i \frac{b_\mathrm{i}}{\nu_{0\mathrm{i}}^2 - \nu^2}\,. \tag{27.8}$$

This *dispersion formula*[4] is well fulfilled for gases and vapors, of course apart from the regions of their eigenfrequencies ν_0. For liquids and solids, it is however nothing more than a useful interpolation formula. Table 27.1 contains a numerical example for rock salt, NaCl.

The discrepancies between the calculated and the observed values is never greater than 5 units in the third place after the decimal point, although only a single eigenfrequency $\nu_0 = 2.85 \cdot 10^{15}$ Hz was employed. It corresponds to the wavelength $\lambda_0 = 0.105\,\mu$m. It could

Table 27.1 The dispersion of NaCl between $\lambda = 0.3\,\mu$m and $5\,\mu$m (Fig. 27.1)
($N_V = 2.28 \cdot 10^{28}$ ion pairs/m^3, $b = 4$, $i = 1$, $\nu_0 = 2.85 \cdot 10^{15}$ Hz $> \nu$)

λ in μm	0.3	0.4	0.5	0.7	1	2	5
n measured	1.607	1.568	1.552	1.539	1.532	1.527	1.519
n computed from Eq. (27.8)	1.610	1.567	1.550	1.535	1.528	1.522	1.521

[4] The quotient $\frac{n^2-1}{n^2+2} = R'$ is known as the *refraction*.

be called the "center of gravity" of the k curve in the ultraviolet (Fig. 27.1). Of course, by using $i = 3$ or 4, we could improve the agreement between the calculation and the measurements even further; but that would be rather unproductive.

27.9 Refractive Indices for X-rays

As an additional test of the dispersion formula (Eq. 27.8), we determine the refractive index for X-rays. With X-rays, the chemical bonding of atoms into molecules plays practically no role (Sect. 27.12). N_{V} in Eq. (27.8) thus refers to the number density of the *atoms*, i.e.

$$N_{\mathrm{V}} = \frac{N_{\mathrm{A}}\varrho}{M_{\mathrm{n}}}$$

(N_{A} is the AVOGADRO constant $= 6.022 \cdot 10^{23}\,\mathrm{mol}^{-1}$, and M_{n} is the molar mass, M/n).

Furthermore, b is the number of all the electrons in an atom, $b = \Sigma b_i$. This number is the same as the atomic number Z, which can be expressed in terms of the molar mass $M_{\mathrm{n}} = A_{\mathrm{r}}$ kg/kmol (see Sect. 26.7); for atoms whose molar mass is not too large, $Z = 0.5\,A_{\mathrm{r}}$ (Eq. (26.21)). Therefore,

$$N_{\mathrm{V}}b = \frac{\text{Number of electrons}}{\text{Volume}} = 0.5 \cdot 6.022 \cdot 10^{26}\,\frac{1}{\mathrm{kg}}\,\varrho\,.$$

In addition, we consider only one resonance frequency ν_0; and finally, for hard X-rays, $\nu_0 \ll \nu$. Using $\nu = c/\lambda$, we obtain from Eq. (27.8)

$$\frac{n^2-1}{n^2+2} = 26.9\,\frac{\mathrm{m}^3}{\mathrm{s}^2}\,\frac{N_{\mathrm{V}}b}{(-\nu^2)} = -0.81 \cdot 10^{28}\,\frac{\mathrm{m}^3}{\mathrm{s}^2\mathrm{kg}}\,\varrho\left(\frac{\lambda}{c}\right)^2. \qquad (27.9)$$

Numerical example: $\varrho = 10^4\,\mathrm{kg/m}^3$ and $\lambda = 10^{-10}$ m, refractive index $n = 0.999986$, thus somewhat less than 1. This is confirmed by observations; see Sect. 27.2.

The dispersion equation (27.8) thus encompasses the whole spectrum from the infrared up to the X-ray region. It also functions in the region of the longest waves; however there, we must take into account not only the secondary radiation from electrons, but also that from ions or from still larger objects.

27.10 The Refractive Index and the Number Density. Light Drag

We have interpreted refraction in terms of the secondary radiation from the irradiated molecules, and thus obtained the dispersion for-

Table 27.2 The electric polarizability α of single molecules in AC fields at the frequency of yellow light, $\nu = 5.1 \cdot 10^{14}$ Hz, calculated using Eqns. (27.3) and (27.4). Compare the values determined here for α with those found for low frequencies in Table 13.5.

Substance	Mass density ϱ in $\frac{\text{kg}}{\text{m}^3}$	Number density of molecules N_V in m^{-3}	Measured refractive index n at $\lambda = 589\,\text{nm}$	Refraction $R' = \dfrac{n^2-1}{n^2+2}$	Molecular polarizability α in $\frac{\text{A s m}}{\text{V/m}}$
O_2 liquid, −183°C	1130	$2.14 \cdot 10^{28}$	1.222	$1.41 \cdot 10^{-1}$	$1.77 \cdot 10^{-40}$
O_2 gas, 0°C and $1.013 \cdot 10^5$ Pa	1.43	$2.69 \cdot 10^{25}$	1.00027	$1.82 \cdot 10^{-4}$	$1.78 \cdot 10^{-40}$
Water, liquid	1000	$3.36 \cdot 10^{28}$	1.334	$2.06 \cdot 10^{-1}$	$1.64 \cdot 10^{-40}$
Water vapor, 0°C, referred to $1.013 \cdot 10^5$ Pa	0.805	$2.69 \cdot 10^{25}$	1.000255	$1.7 \cdot 10^{-4}$	$1.68 \cdot 10^{-40}$

mula (Eq. (27.8)). It provides two kinds of information: First, it describes the dependence of the refractive index n on the light frequency ν. That was shown in Sects. 27.7–27.9. Second, it describes the influence of N_V, the number density of the molecules, on the value of the refractive index. We want to consider this point in more detail. To that end, we take ν to be constant, and thus consider observations using monochromatic light.

For gases and dilute solutions[5], n is nearly equal to 1, and thus the refraction $R' = (n^2 - 1)/(n^2 + 2)$ is nearly equal to $\frac{2}{3}(n-1)$. Then, instead of Eq. (27.8), we have

$$(n-1) = \text{const} \cdot N_\text{V}\,, \tag{27.10}$$

i.e. in gases, $(n - 1)$ is proportional to the number density, and in solutions, it is proportional to the concentration. Or, expressed differently, each molecule makes its contribution to the refractive index independently of all the other molecules. This relation between the refractive index and the gas density is well suited for demonstration experiments. An experiment for a teaching laboratory was illustrated in Fig. 20.31.

The independent contributions of the individual molecules to refraction remains valid even for the gas – liquid transition, that is in spite of a change in density of around 1:1000. In Table 27.2, we have listed some numerical examples, both for the refraction R' as well as for the *electric polarizability* α which is derived from it. Both quantities are thus to a large extent *independent of the aggregate state* (solid, liquid, gaseous) and of *chemical bonding*.

A strange effect, called the "*drag* of light", was predicted in 1818 by A. Fresnel and found in 1851 by A.H.L. Fizeau: The propagation of a light wave is influenced by the motion of the molecules. A liquid

[5] In solutions, $n = \dfrac{n_\text{solution}}{n_\text{solvent}}$, and it thus represents the refractive index that is due solely to the dissolved molecules.

which is moving in the direction of the light propagation at the velocity u has a smaller refractive index for an observer at rest outside the liquid, i.e. the velocity of the light is greater than for the same liquid at rest. The velocity u of the liquid thus changes the phase velocity c/n of light within the liquid, but not by its full magnitude $\pm u$; instead, by approximately $\pm u\,(1-1/n^2)$. (The experimental arrangement is the same as shown in Fig. 20.31, except that both light beams pass through chambers filled with water flowing in opposite directions.)

We can understand the 'light drag' qualitatively from Eq. (27.8): The number density N_V of the molecules acted upon by the light increases when the liquid is flowing towards the light source, and *vice versa*. Quantitatively, the drag is calculated from the LORENTZ transformations in the special theory of relativity.

27.11 Curved Light Beams

The refractive index which holds for a particular monochromatic radiation depends on the number density N_V of the active molecules (Eq. (27.8)). This density can be varied continuously within a certain volume, thus producing a gradient in the refractive index. In such a volume, *curved* light beams can be observed, as seen for example in Fig. 27.14. The boundaries of curved beams as well as their axes are drawn as curved light rays. In general, the radius of curvature of a ray changes along its path. At each position x, we have

$$r = \frac{n}{\mathrm{d}n/\mathrm{d}r} \tag{27.11}$$

(derivation in Fig. 27.15).

Here, $\mathrm{d}n/\mathrm{d}r$ is the gradient of the refractive index at the position x in *a direction perpendicular to the beam.*

Experimentally, gradients in the refractive index can be produced by concentration gradients in solutions. The most suitable method is

Video 27.1:
"Curved light beams"
http://tiny.cc/wfggoy
A laser is used as light source, so that there is no fanning-out of the beam due to dispersion. To produce the gradient in the refractive index, seven layers of sugar solution, each 1 cm thick and with decreasing concentration, were laid on each other. While each layer was being poured, a thin cork sheet was used to keep mixing to a minimum.

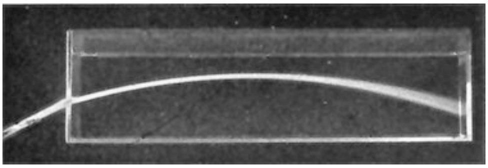


Figure 27.14 A curved light beam in a liquid with a vertical, approximately linear gradient in its refractive index. The fanning-out of the beam at the right is due to dispersion: The path of the waves with the shortest wavelengths is the most strongly curved (this is also a model experiment for the appearance of the "green flash" (see below)) **(Video 27.1)**.

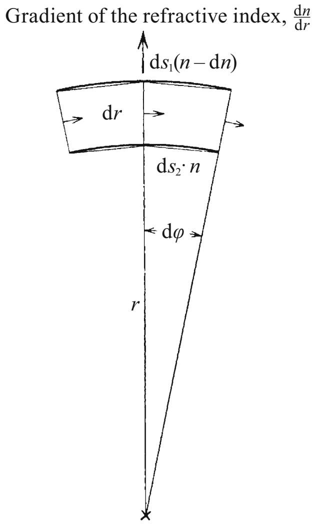

Gradient of the refractive index, $\frac{\mathrm{d}n}{\mathrm{d}r}$
$\mathrm{d}s_1(n-\mathrm{d}n)$
$\mathrm{d}r$
$\mathrm{d}s_2 \cdot n$
$\mathrm{d}\varphi$
r

Figure 27.15 The derivation of Eq. (27.11). The three arrows indicate the tipping of the wave crests at their ends. The optical path lengths are given according to Eq. (16.5) by $\mathrm{d}s_1 \cdot (n - \mathrm{d}n) = \mathrm{d}s_2 \cdot n$. Furthermore, from the sketch, we can see the geometrical relation $\mathrm{d}s_1 = \mathrm{d}\varphi \cdot (r + \mathrm{d}r)$, $\mathrm{d}s_2 = \mathrm{d}\varphi \cdot r$. Combining the three equations leads to Eq. (27.11).

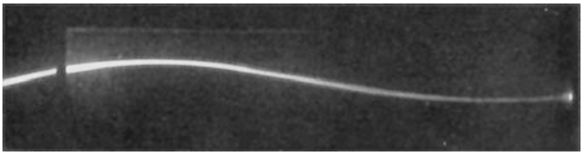


Figure 27.16 A light beam with a wavy shape. The refractive index has its maximum value at the center of the tank. Below is a saturated alum solution with a density of 1.04 g/cm^3. Above it is a glycerine/alcohol mixture, roughly 1:1, of density 1.01 g/cm^3. At the top is water with ca. 10 % alcohol, density 0.98 g/cm^3. All the solutions contain quinine sulfate and sulfuric acid, and their boundaries were smeared out by several hours of diffusion. The recipe is due to R.W. Wood.[C27.7]

[C27.7] C27.7. See R.W. Wood, *Physical Optics* (McMillan Publishers, NY, 3rd edition 1934), p. 90.

to use two liquids which are completely miscible, and to add them in layers with appropriately chosen compositions. The boundary surfaces which are initially present between the layers are rapidly smeared out by diffusion. In this way, an approximately linear gradient in the refractive index was produced in Fig. 27.14. At the bottom of the tank is pure carbon disulfide ($n = 1.63$); at its top, pure benzene ($n = 1.50$), and the transition between the two consists of around 10 layers, each 1 cm thick. The light beam is most strongly curved at its highest point, i.e. its radius of curvature r has its minimum value there. This corresponds to Eq. (27.11): At the top of the curve, the gradient of the refractive index in the direction *perpendicular* to the direction of the light beam is maximal.

In Fig. 27.16, the gradient in the refractive index is also vertical, but it changes its sign in the center of the tank. In this way, we can demonstrate a light beam with a wavy shape.

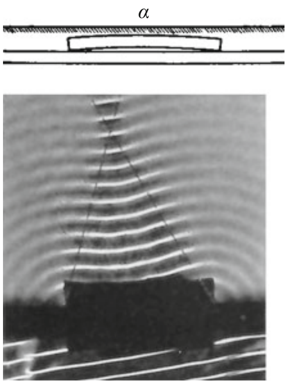


Figure 27.17 The lens action of a cylindrically-symmetric gradient in the refractive index (‘snapshot’). Top: The cross-section of the arched rectangular piece of sheet metal under the water

Radially-symmetric gradients in the refractive index, some with cylindrical symmetry and some with spherical symmetry, play an important role in the eyes of animals. The most prominent example are the compound facette eyes of insects in their various forms. However, even in the lenses of the eyes of vertebrates, there are also gradients in the refractive index in combination with the curvature of the lens surfaces. Strictly considered, in a sketch of the human eye, one should draw curved light rays within the eye’s lens.

“Strictly considered, in a sketch of the human eye, one should draw curved light rays within the eye’s lens”.

Owing to its importance, we want to translate image formation with curved light rays into the wave picture. To this end, we show a model experiment with water waves in Fig. 27.17. We start from Fig. 27.11 b and set a flat, rectangular, cylindrically-arched piece of sheet metal under the surface of the water, between the two sides of the slit. Its cross-section is sketched in Fig. 27.17 (top). Its long axis is set perpendicular to the slit; this gives a shallow-water area of rectangular shape and variable depth. The depth of the water is smallest in the center at α, and greatest at the side edges. As a result, the waves propagate more slowly in the center than at the edges (Vol. 1, Eq. (12.36)). They are convergent when they leave the rectangular area and are focussed at an image point (Fig. 27.17 (bottom)).

Gradients in the refractive index with *spherical symmetry* play an important role in astronomical observations. We mention just one example. The density of earth’s atmosphere decreases on going upwards from the surface. A light beam which is incident tangential to the earth’s surface arrives at the eye of an observer on a curved path. The sun, when it seems to just touch the horizon, has actually already set; the “atmospheric ray refraction” causes it to appear too high by 32 minutes of arc. This means that a surprising effect can occur during an eclipse of the moon: We can see the sun and the darkened moon opposite to each other, both just above the horizon, at the same time.

At sunset, especially at sea, one can sometimes see the last part of the disappearing disk of the sun flash brightly with a green color. This phenomenon, known as the “green flash”, can be explained by the stronger curvature of the optical path for short-wavelength light (Fig. 27.14) (and *not* by a contrast effect in the eye of the observer).

The *gravitational field* of the earth is only indirectly involved in the phenomenon of atmospheric ray refraction. Together with the thermal molecular motions, it produces the density gradient of the gas molecules and thus the gradient of the refractive index in the atmosphere.

Surprisingly, however, gravitational fields alone can also produce a gradient in the refractive index even in empty space, without any interacting molecules. The light from the stars is subject to a *deflection* of about 1.75 seconds of arc near the disk of the sun (it is visible only during total solar eclipses).

Half of this deflection, that is 0.88″, was already explained in 1801 by J. G. v. SOLDNER: Light rays in gravitational fields act like projectiles emitted from the stars at a velocity of $u = 3{\cdot}10^8$ m/s. They propagate along hyperbolic paths. The other half of the observed deflection was predicted by A. EINSTEIN with his general theory of relativity.[C27.8]

27.12 The Qualitative Interpretation of Absorption

We begin by looking back at Figs. 27.1 and 27.2. The extinction spectra consist in general of a number of individual bell-shaped bands. As a rule, they are not completely separated from each other, and often, single narrow bands merge together to give broad "unresolved" bands.

We note for a preliminary overview: In the region of hard X-rays, the extinction spectra are determined only by the *atoms*. They consist of the sum of the extinction spectra of all the atoms involved. Chemical bonding and the aggregate state have practically no influence. Our conclusion: The extinction of the radiation occurs in the innermost electronic shells of the atoms, which are protected from the influence of their surroundings.

In the region of soft X-rays, chemical bonding begins to make itself felt, along with the aggregate state: Crystals exhibit new bands which are lacking in the spectra of individual molecules. We conclude that now, the middle and outer electronic shells of the atoms play an important role; they are no longer impervious to influences from the surroundings of the atoms.

In all of the remaining spectral regions, from the ultraviolet through the visible to the infrared, the extinction spectra of the atoms depend to a large extent on the aggregate state of the material. Furthermore, bonding of the atoms into molecules gives rise to new, additional bands. Conclusion: Here, the extinction is produced by processes in the outermost electronic shells, which are also responsible for chemical bonding, formation of the liquid phase and of crystal structure.

C27.8. EINSTEIN also initially (1911) obtained a similar value to that of v. SOLDNER due to the relativistic time dilation. However, after his theory was completed in 1915, now containing also a spatial dilation, he predicted an overall light deflection of 1.75 seconds of arc. This value was then confirmed by observations during the solar eclipse of 1919. See e.g. H. v. Klueber, "The determination of Einstein's light deflection in the gravitational field of the sun", *Vistas in Astronomy*, Vol. III, p. 47 (1960).

Dispersion curves can be explained in terms of forced oscillations: We assumed that electrical resonators were present in the interior of the molecules. Their eigenfrequencies ν_0 coincide with the frequencies of the maxima of the absorption bands. Given these circumstances, we are led almost inevitably to an explanation of the *absorption process*: The *damping* of the resonators consumes a part of the energy of the incident light and converts it into other forms of energy, e.g. into heat. Here, again, we offer a simple analogous example from mechanics:

A glass filled with white wine makes a clear ringing sound when clinked with another glass. The glasses and their contents are subject to oscillations (standing waves). These are formed by the superposition of travelling waves which are constantly reflected from the walls. A glass filled with champagne, in contrast, cannot be made to ring by clinking it with another glass; champagne contains bubbles of gas. They themselves also act as resonators: They are excited to forced oscillations by the waves. Their damping consumes the energy of the waves.

"A glass filled with white wine makes a clear ringing sound when clinked with another glass . . . But a glass filled with champagne cannot be made to ring by clinking".

27.13 A Quantitative Treatment of Absorption

In Sect. 27.12, we gave a qualitative explanation of absorption. It is supported by the shape of single absorption bands, i.e. those that are well separated from neighboring bands. They often exhibit a noticeable similarity to the energy-resonance curves of forced oscillations (Vol. 1, Fig. 11.44). There, the ordinate is a measure of the kinetic energy stored in the resonator, or the average power consumed by the damping, $\overline{\dot{W}}_\nu$.

The quantitative treatment is closely related to that in Sect. 26.5. We again assume that the electric resonators are dipoles. We initially make no assumptions about their nature. Their number density is N'_V. The incident light is again taken to be in the form of a collimated beam. The absorbing substance is in a *dilute solution* and the solvent has a refractive index of n.

In a segment of the beam of length Δx and cross-sectional area A, there are $N'_\text{V} A \Delta x$ damped *resonators*. They give rise to an absorption constant

$$K = \frac{\Delta \overline{\dot{W}}_\nu}{\overline{\dot{W}}_\text{p}} \cdot \frac{1}{\Delta x}\,. \qquad \text{(Defining equation (25.1))}$$

Here, $\Delta \overline{\dot{W}}_\nu$ is the power consumed by the resonators, and

$$\overline{\dot{W}}_\text{p} = n \frac{\varepsilon_0}{2} E_0^2 c A \qquad (27.12)$$

is the power of the radiation which passes through the cross-section A and excites the resonators there.C27.9 $\Delta\overline{\dot{W}}_\nu$ is the sum of the power consumed by all of the damped resonators together. Each single one of them consumes the power

C27.9. Derivation: In vacuum, we have

$$\overline{\dot{W}}_\mathrm{p} = \frac{1}{2\mu_0} E_0 B_0 A = \overline{S} A\,,$$

where $\overline{S}$ is the time averaged value of the POYNTING vector (Comment C12.4). In a dielectric material (refractive index n), we have

$$B_0 = E_0/(c/n)$$

and thus

$$\overline{\dot{W}}_\mathrm{p} = \frac{n}{2\mu_0 c} E_0^2 A\,,$$

and with $c = (\varepsilon_0\mu_0)^{-1/2}$ (Eq. (8.7)), Eq. (27.12) follows.

$$\overline{\dot{W}}_\nu = -4\pi H \cdot \overline{W}_\mathrm{kin}\,. \tag{27.13}$$

H is the halfwidth of the energy-resonance curve, as defined in Vol. 1, Fig. 11.42, and $\overline{W}_\mathrm{kin}$ is the average value of the kinetic energy stored in a resonator when it is oscillating at the frequency ν.

Derivation: The amplitude $\alpha(t)$ of the free damped oscillation follows an exponential law:

$$\alpha(t) = \alpha_0 e^{-\Lambda t/T}\,, \tag{27.14}$$

where $T = \nu_0^{-1}$ is the period of the harmonic oscillator (Vol. 1, Sect. 11.10). For the energy, it follows that

$$W_{\nu_0} = W_0 e^{-2\Lambda t/T} \quad \text{and} \quad \overline{\dot{W}}_{\nu_0} = -\frac{2\Lambda}{T}\overline{W}_{\nu_0} \tag{27.15}$$

(the power here is the value averaged over a period T, since, precisely considered, the energy consumption is 'pulsed' due to the dependence of the frictional force that causes the damping on the momentary velocity). Together with Eq. (11.2), and using $\overline{W}_{\nu_0} = 2\overline{W}_{\mathrm{kin},\nu_0}$, we finally obtain (**Exercise 27.1**):

$$\overline{\dot{W}}_{\nu_0} = -2\pi H \overline{W}_{\nu_0} = -4\pi H \overline{W}_{\mathrm{kin},\nu_0}\,. \tag{27.16}$$

In the steady state, this power must be continually replaced. This result also holds when the oscillator is excited at a frequency above its resonance frequency ν_0 (not shown here). This leads to Eq. (27.13).

All of the resonators contained in the volume $A\Delta x$ together consume the power denoted by $\Delta\overline{\dot{W}}_\nu$:

$$\Delta\overline{\dot{W}}_\nu = N'_\mathrm{V} A \Delta x 4\pi H \overline{W}_\mathrm{kin}\,.$$

With Eq. (26.1), we find the time-averaged value of the kinetic energy of one oscillator which is vibrating at the frequency ν (**Exercise 27.2**):

$$\overline{W}_\mathrm{kin} = \frac{1}{4} m(\omega l_0)^2 = \left(\frac{1}{4\pi}\right)^2 \frac{e^2 E_\mathrm{w}^2}{m} \cdot \frac{\nu^2}{(\nu_0^2 - \nu^2)^2 + \left(\dfrac{\Lambda}{\pi}\right)^2 \cdot \nu_0^2 \nu^2}\,. \tag{27.17}$$

The amplitude F_0 of the exciting force was not set here to eE_0, but rather to eE_w. E_w is the exciting amplitude of the light for a single resonator. In materials with a refractive index $n > 1$ (liquids and crystals), it is larger than the field-strength amplitude in vacuum, E_0.

Equation (13.32) holds here ($n = \sqrt{\varepsilon}$):

$$E_\mathrm{w} = \frac{E_0}{3}(n^2 + 2)\,.$$

Combining these equations leads to the absorption constant

$$K = \frac{N'_\mathrm{V} e^2}{2\pi c \varepsilon_0 m} \cdot \frac{(n^2+2)^2}{9n} \cdot \frac{H \cdot \nu^2}{(\nu_0^2 - \nu^2)^2 + \left(\dfrac{\Lambda}{\pi}\right)^2 \cdot \nu_0^2 \nu^2}\,. \qquad (27.18)$$

This equation gives information about the *shapes* of the optical absorption curves (Sect. 27.14). In addition, it offers us the possibility of determining the number density N_V of the resonators from an optical measurement: It leads to a quantitative analysis method for absorption spectra, as discussed in Sect. 27.15.

In both cases, we must keep an essential point in mind: In deriving Eq. (27.18), we did not take a mutual influence of the resonators on each other into account. For this reason, this equation and its various rearranged forms (Eq. (27.19)) can be applied only to the limiting case of dilute solutions or gases of moderate density (see the footnote in Sect. 27.10).

27.14 The Shapes of Absorption Bands

For a given solution, the first two fractional factors in Eq. (27.18) contain only constants. (The weak dependence of the refractive index n on the frequency ν can be neglected within the width of a band). Then we can choose the maximum value K_max arbitrarily by fixing N'_V and obtain the shape of the absorption band from the fractional expression on the right.

Figure 27.18 shows two examples. The upper part of the figure refers to a solid solution of potassium atoms in a KBr crystal. In this system, a small fraction of the Br ions in the lattice has been replaced by electrons[6]. Their number density is N_V. This gives rise to new absorbing centers for which the name *color centers* (F centers) has become established. The lower part of the figure shows data for a vapor solution of mercury in compressed hydrogen. It contains about one Hg atom for each $6 \cdot 10^6$ $\mathrm{H_2}$ molecules.

Note the different scales on the abscissas in Fig. 27.18. In the upper graph, the absorption curve is a broad *band* with a halfwidth of $H = 1.21 \cdot 10^{14}$ Hz (Q-factor $\nu_0/H = 3.8$). The lower graph, in contrast, shows a *spectral line* whose width is determined by thermal collisions

[6] F centers. Considered chemically, a $\mathrm{K^+}$ ion together with an electron forms a neutral potassium atom. H. Pick, “*Struktur von Störstellen in Alkalihalogenidkristallen*”, Springer Tracts in Modern Physics (Springer-Verlag Berlin, Vol. 38, p. 1 (1965)).

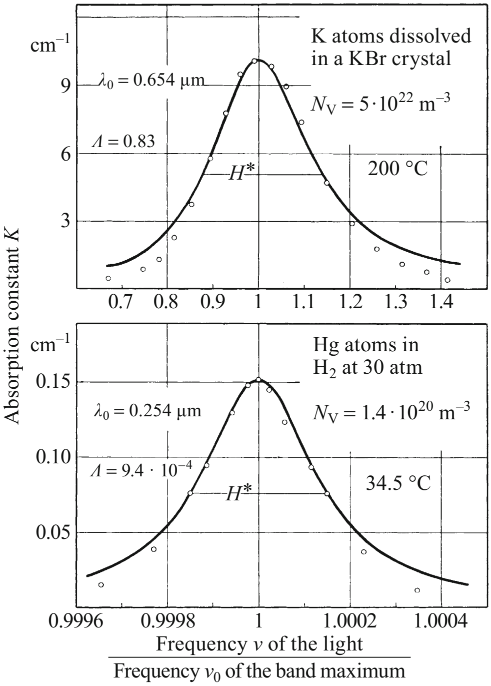


Figure 27.18 The representation of optical absorption bands as energy-resonance curves (lower part from measurements by G. JOOS). The calculated curves are fitted to the measurements at the band maxima. ($H^* = H/\nu_0$, 1 atm = $1.013 \cdot 10^5$ Pa ≈ atmospheric pressure)

(thermal broadening). It has $H = 3.54 \cdot 10^{11}$ Hz ($\nu_0/H = 3.3 \cdot 10^3$). In both examples, the calculated curves agree quite satisfactorily with the measurements. Thus, the basic assumption underlying this calculation, that of *exponentially-damped* resonators, combined with the fitting of the number density N_V to the observed maximum, yields a useful model for the real situation. This is, however, not the case for every absorption band. The systematic deviations between calculated and measured bands are in most cases considerably greater than in Fig. 27.18. In such cases, one can consider *exponentially*-damped resonators (e.g. electrons which are quasi-elastically bound to positive charges, dipoles) only as an approximation. This model has however the advantage that it is intuitively understandable.

27.15 Quantitative Analysis Using Absorption Spectra

Setting $\nu = \nu_0$ in Eq. (27.18), we obtain the absorption constant $K_{\max}$ at the band maximum. At the same time, we can solve for N'_V, remove the logarithmic decrement by using the approximate relation $\Lambda = \pi H/\nu_0$ (for $\Lambda \leq 1$), and find

$$N'_{\mathrm{V}} = \frac{2\pi c\varepsilon_0 m}{e^2}\,\frac{9n}{(n^2+2)^2}\,K_{\max}\cdot H\,. \tag{27.19}$$

We now compare the number density N'_{V} of the resonators as determined from the absorption band with the overall number density N_{V} of absorbing centers ("molecules"). To do this, we define the ratio

$$\frac{N'_{\mathrm{V}}}{N_{\mathrm{V}}} = \frac{\text{Number of resonators}}{\text{Number of molecules}} = f\,.$$

This number f is called the "oscillator strength". In Eq. (27.19), only constants precede the product $K_{\max}H$. We thus obtain for the number density of the absorbing molecules:

$$N_{\mathrm{V}} = \text{const}\cdot K_{\max}\cdot H \tag{27.20}$$

with

$$\text{const} = \frac{2\pi c\varepsilon_0 m}{f\cdot e^2}\cdot\frac{9n}{(n^2+2)^2} \tag{27.21}$$

(c = velocity of light; ε_0 = electric field constant; m = mass and e = charge of the electron; n is the refractive index of the solvent at the frequency of the band maximum (KBr in the example in Fig. 27.18, top); f = oscillator strength).

This constant can thus be computed from physical constants, from the refractive index n of the solvent, and from the oscillator strength f. Equation (27.20) offers the possibility of determining N_{V}, the number density of the absorbing molecules, from optical measurements. This procedure, corresponding to the derivations of Eqns. (27.19) and (27.20), is limited to the case that the resonators are independent and do not mutually influence each other (BEER's law, see Sect. 25.4). In particular cases, one can set $f = 1$. Usually, however, the constant must be empirically evaluated by a calibration measurement with a relatively large number density N_{V} which can be determined chemically.

Spectral analysis using optical absorption is superior to chemical analysis in terms of sensitivity. We estimate the orders of magnitude: The constant in Eq. (27.20) has a magnitude of about $6\cdot 10^5\ \mathrm{s/m^2}$. At a layer thickness of 10 cm, absorption constants down to $K = 0.01/\mathrm{cm}$ can be measured ($e^{-0.1} = 0.9$). The decisive factor is now the halfwidth H. For solid and liquid solvents, H is seldom less than 10^{14} Hz. With these values for $K_{\max}$ and H, we can optically determine number densities down to $N_{\mathrm{V}} = 10^{20}/\mathrm{m}^3 = 10^{14}/\mathrm{cm}^3$. In solid and liquid substances, the number density of the molecules is of the order of $10^{28}/\mathrm{m}^3$. Therefore, we can optically detect one dissolved molecule among 10^8 molecules of a solid or liquid solvent.

In gases and vapors, the halfwidth H is considerably smaller; values of 10^{10} Hz are not unusual[7]. Then the absorption in a 10 cm thick sample suffices to detect molecules with a number density of about $10^{16}/\text{m}^3$. Such a value of the number density corresponds to a vapor pressure of the order of 10^{-3} Pa.

At room temperature, mercury has a saturation vapor pressure of 0.16 Pa. In poorly-ventilated laboratories where mercury is used without sufficient protection, there can be just as many mercury vapor molecules in each $1\,\text{m}^3$ of air as in a mercury droplet with a volume of $1\,\text{mm}^3$. Optically, we can already detect 1 % of this amount. For the absorption measurements, one uses the wavelength $\lambda = 0.2537\,\mu\text{m}$ (cf. 13th edition of "*Optik und Atomphysik*", Fig. 14.19).

Absorption spectral analysis has also been employed in liquid and solid substances, for example to detect and measure the concentration of vitamins,[C27.10] and for the physical investigation of photochemical reactions in crystals.[C27.11]

C27.10. R.W. Pohl, *Naturwissenschaften* **15**, 433 (1927); A. Windaus, Nobel Lectures, Chemistry, 1927 (Elsevier Amsterdam, 1966), p. 105.

C27.11. A. Smakula (Dr. rer. nat., Göttingen, 1927), *Zeitschrift für Physik* **59**, 603 (1930); see also the footnote in Sect. 27.14.

27.16 The Properties of Optically-Effective Resonators

We have seen that the classical interpretation of dispersion and absorption in terms of the forced oscillations of resonators is able to reproduce the observations to a good approximation. We therefore want to supplement this model by giving some information on the various types of resonators that may play a role.

Light, just like an alternating electric field, gives rise to *influence* in molecules; they are electrically "deformed" or "polarized": The centers of gravity of their positive and negative charges are shifted relative to one another. This periodic variation of the charge distribution is usually represented by the schematic picture of an oscillating dipole. Two elementary charges of opposite sign are assumed to be at its ends, that is $\pm 1.6 \cdot 10^{-19}$ A s.

The mass of the molecule can be distributed in various ways over the two charge carriers. In a *limiting case*, the negative charge is assumed to carry only the small mass of one electron, i.e. $9 \cdot 10^{-31}$ kg, and all of the remaining large mass of the molecule is attached to the positive charge. Then the molecule remains practically at rest as a positive ion, and the dipole is created only through the oscillations of the electron around its rest position[8]. One refers for short to a *quasi-elastically bound* electron. This model was found in the

[7] Note that the measurements on mercury in Fig. 27.18 were carried out at a pressure of $30 \cdot 10^5$ Pa.

[8] The eigenfrequency ν_0 of such a dipole (resonator or oscillator) corresponds in quantum mechanics to the frequency $\nu_0 = \Delta W/h$ when the energy of the molecule changes by ΔW.

Table 27.3 Velocity of longitudinal sound waves, lattice constant and frequency of the residual-ray bands of some crystals

Crystal	Sound velocity in m/s	Spacing D of neighboring lattice components (positive alkali ion and negative halogen ion)	Frequency of the residual-ray band	
			Calculated from Eq. (27.22)	Observed
NaCl	$3.3 \cdot 10^3$	$2.81 \cdot 10^{-10}$ m	$5.9 \cdot 10^{12}$ Hz	$5.8 \cdot 10^{12}$ Hz
KCl	3.1	3.14	4.9	4.7
KBr	2.3	3.29	3.5	3.6
KI	1.95	3.52	2.8	2.7

preceding sections to give good results both for visible light as well as for ultraviolet light and X-rays.

The situation is different in the infrared spectral region: There, we encountered the absorption bands belonging to the *residual rays*. They were observed for cubic ionic crystals (Fig. 27.9). A platelet made from one of these crystals can be thinned at most to the spacing D of two neighboring lattice components, for example a Na^+ and a Cl^- ion in NaCl. Such a (limiting-case) platelet of thickness D would have a mechanical eigenfrequency of

$$\nu_0 = \frac{u}{2D}\,. \tag{27.22}$$

Here, u is the velocity of longitudinal *sound waves* in the crystal. This *mechanical* frequency agrees with the *optical* frequency of the residual rays. This is shown by the numbers[9] in Table 27.3.

Thus, for the case of 'residual rays', we can calculate an *optical* frequency from data obtained by *non*-optical measurements. This is the fundamental importance of this relation, which was discovered in 1908 by E. Madelung.

This fact at the same time leads to information about the type of resonators that give rise to residual rays: Both of the elementary charges are bound to the large masses of *ions*. These ions, e.g. Na^+ and Cl^-, vibrate oppositely to each other and thus form an oscillating dipole. Here, the picture of a dipole is already *more* than just a model scheme.

In the simplest ionic crystals, of the type NaCl, the molecules have lost every trace of an individual existence. This is however a limiting case. In many other types of crystals, whole molecules or parts of them maintain their own identities. In such independent molecules, and also in molecular crystals, pairwise oppositely-charged molecular components can form dipoles and can absorb infrared light

[9] The oscillation period $T = 1/\nu$ for the fundamental mechanical vibration of a rod is $= 2D/u$. This means that a longitudinal elastic perturbation passes along the entire length D of the rod twice during the time T, namely going forward and then returning. The velocity of sound within a solid body is nearly always quoted as the special case of its value along the length of a rod, without further comment (cf. Vol. 1, Fig. 12.42 and Eq. (11.5)). In Eq. (27.22), however, a valid average value for a three-dimensional body must be used.

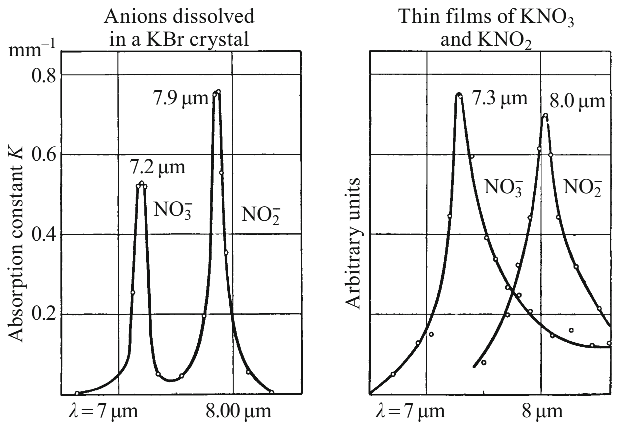


Figure 27.19 Absorption spectra of NO_3^- and NO_2^- ions. For the right-hand graph, thin crystalline layers of KNO_3 and KNO_2 were employed. For the left-hand graph, a solid solution of the ions in a KBr crystal was used. The concentration was ca. 0.1 % in the melt from which the crystal was prepared. In the crystal, the concentration is roughly ten times lower than in the melt.

through forced oscillations. Two of many possible examples can be found in Fig. 27.19. The two parts of the figure show absorption bands belonging to the NO_3 group and to the NO_2 group. They lie at around 7.2 and 8.0 µm. The right-hand image refers to KNO_3 and KNO_2 crystals, and the left-hand image to a solid solution of these compounds in a KBr crystal. In this second case, mixed crystals were grown, in which individual Br^- ions were replaced in part by NO_2^- and in part by NO_3^- ions. In spite of the different crystal structures, the absorption bands of NO_3 and NO_2 are in both cases at practically the same positions. Thus, the absorption of infrared radiation leads us to the recognition of *inner* oscillation frequencies which are characteristic of the individual molecules. We must however be careful to avoid an error: Large molecules built up from many components can have many eigenfrequencies (compare Vol. 1, Sect. 11.5), but only some fraction of those frequencies are associated with the oscillations of electrically-charged molecular components. Only *those* oscillations can be effective in producing absorption bands (only they are "optically active"). The optical detection of the remaining frequencies must be carried out by a different method (described in the 13th edition of "*Optik und Atomphysik*", Chap. 15, 'Raman scattering', or e.g. www.tsi.com/basics-of-raman-spectroscopy/).

The *permanent* electric dipole moments of *polar* molecules have no importance for the absorption and dispersion in the optical spectral regions. Their role begins only in the range of electric waves (radio, microwaves). There, liquids with dipolar molecules can exhibit strong absorption and large refractive indices. A well-known example is water (cf. Sect. 13.11).

27.17 The Mechanism of Light Absorption in Metals

The absorption spectra of metals show a special feature: In all the *non-metallic* substances, the bands from "bound" electrons are followed by an *absorption-free region* (Fig. 27.1). Only at longer wavelengths, in the infrared region, do we begin to see absorption due to ions. In metals, in contrast, an additional absorption is observed, beginning in the ultraviolet, which initially increases in strength with increasing wavelength. Usually, it overlaps with the longest-wavelength bands that result from bound electrons (Fig. 27.6). There is thus no absorption-free region in the spectrum, and the absorption constant in the infrared attains values of the order of $10^5\,\text{mm}^{-1}$.

This additional absorption, which is lacking in all other materials but is seen in metals, is due to their electrical *conductivity* σ; it thus has its origin from "free" or "conduction" electrons. At $\lambda > 10\,\mu\text{m}$, practically the only absorption is that due to free electrons. There, just as in the range of electric waves (radio, microwaves), we can calculate the absorption from the conductivity σ. The magnetic field of the *penetrating* waves generates eddy currents which convert the energy of the waves into heat. The relations derived for electromagnetic waves apply, namely

$$n = k = \sqrt{\frac{1}{4\pi\varepsilon_0}\cdot\frac{\sigma}{\nu}} = 5.47\sqrt{\text{ohm}}\cdot\sqrt{\sigma\cdot\lambda} \tag{27.23}$$

and

$$K = \sqrt{\frac{4\pi}{\varepsilon_0 c}\cdot\frac{\sigma}{\lambda}} = 68.8\sqrt{\text{ohm}}\cdot\sqrt{\frac{\sigma}{\lambda}} \tag{27.24}$$

(here, n is the refractive index, k the absorption coefficient, defined by Eq. (25.3), K is the absorption constant, defined by Eq. (25.1), λ is the vacuum wavelength, σ the specific electrical conductivity, and ε_0 is the electric field constant, $8.86\cdot10^{-12}$ A s/V m).

Derivation: The third MAXWELL equation (see Sect. 14.5) states that

$$\text{curl}\,H = j + \dot{D}\,. \tag{27.25}$$

Here, $\dot{D} = \varepsilon\varepsilon_0\dot{E}$ is the displacement current density and $j = \sigma E$ is the conduction current density. For an undamped travelling electromagnetic wave of amplitude E_0, we find from Eq. (25.26) that

$$E = E_0 e^{i\omega(t-zn/c)}, \quad \text{and thus} \quad \dot{E} = i\omega E \quad \text{or} \quad E = -\frac{i\dot{E}}{\omega}\,.$$

Inserting this into Eq. (27.25) yields

$$\text{curl}\,H = \varepsilon_0\dot{E}\left(\varepsilon - \frac{i\sigma}{\varepsilon_0\omega}\right).$$

The expression in parentheses can be regarded as a complex dielectric constant, that is $\varepsilon' = \varepsilon - i\sigma/\varepsilon_0\omega$. It is linked via the MAXWELL relation $(n')^2 = \varepsilon'$ to the complex refractive index $n' = n - ik$. We obtain

$$(n - ik)^2 = \varepsilon - \frac{i\sigma}{\varepsilon_0\omega}\,, \quad \text{i.e.} \quad n^2 - 2nik - k^2 = \varepsilon - \frac{i\sigma}{\varepsilon_0\omega}\,. \qquad (27.26)$$

In metals (and sometimes also in semiconductors), we can neglect the displacement current; that is, we set the real part of the dielectric constant $\varepsilon = 0$. Finally, we can equate the real and the imaginary terms in Eq. (27.26) individually, so that

$$n^2 - k^2 = 0 \quad \text{and} \quad -2nik = -\frac{i\sigma}{\varepsilon_0\omega}\,,$$

and from this, Eq. (27.23) follows.

Numerical examples: For silver, $\sigma = 62 \cdot 10^6\,\Omega^{-1}\cdot\text{m}^{-1}$. At $\lambda = 10\,\mu\text{m}$, $n = k = 136$ and $K = 1.7 \cdot 10^5\,\text{mm}^{-1}$ (compare Fig. 27.6). For mercury, a poorly-conducting metal, the corresponding numbers are: $\sigma = 1.04 \cdot 10^6\,\Omega^{-1}\cdot\text{m}^{-1}$, $n = k = 17.6$, and $K = 2.2 \cdot 10^4\,\text{mm}^{-1}$.

With such large and equal values of n and k, BEER's formula for the reflectivity R can be simplified. Instead of Eq. (25.37), we obtain as a good approximation DRUDE's rule:

$$R = 1 - \frac{2}{k} = 1 - \frac{0.366}{\sqrt{\Omega}\,\sqrt{\sigma\lambda}}\,. \qquad (27.27)$$

In this formula, the absorption coefficient k causes only a (usually small) deviation of the reflectivity from the "ideal" value of 1 (compare Fig. 27.8).

27.18 Dispersion by Free Electrons with Weak Absorption (Plasma Oscillations)

In Sect. 27.8, we derived the dispersion formula (Eq. (27.7)) for spectral regions in which the absorption can be neglected. We can solve Eq. (27.7) for n^2, obtaining

$$n^2 = 1 + \frac{e^2 N_\text{V}}{4\pi^2\varepsilon_0 m} \cdot \frac{1}{\nu_0^2 - (e^2 N_\text{V}/12\pi^2\varepsilon_0 m) - \nu^2}\,. \qquad (27.28)$$

The derivation of Eqns. (27.7) and (27.28) was based on the following considerations:

1. In a neutral molecule, one negative electron and the positively-charged "remainder" of the molecule can undergo mutual quasi-elastic oscillations with an eigenfrequency of ν_0.

2. This oscillatory structure is excited to forced vibrations by an incident light wave.

3. With closely-packed molecules (in liquids and solids), the amplitude of these forced oscillations depends not only on the amplitude of the incident light waves, but also on the electric dipole moments $\boldsymbol{p}$ which are produced in the neighboring molecules under the influence of the light waves.

In this derivation, it was left completely open as to how the quasi-elastic oscillations with an eigenfrequency of ν_0 were produced. One can treat them as a *circular oscillation*, in which the electron orbits on a circular path around the positive charge. The required radial acceleration $\omega_0^2 r$ is produced by the attractive force of the positive charge e (COULOMB's law, Eq. (3.8)). We then find

$$\omega_0^2 r = \frac{e^2}{4\pi\varepsilon_0 r^2 m} \tag{27.29}$$

(ε_0 is the electric field constant $= 8.86\cdot 10^{-12}$ A s/V m; $e = 1.6\cdot 10^{-19}$ A s; m is the mass of the electron $= 9.11\cdot 10^{-31}$ kg).

A circular orbit of radius r requires a volume

$$V = \frac{4}{3}\pi r^3 = \frac{e^2}{3\varepsilon_0 m\omega_0^2}\,. \tag{27.30}$$

This volume V necessary for the circular orbit cannot become larger than $1/N_\text{V}$, i.e. the reciprocal of the number density N_V of the molecules. Then for the largest possible radius, we find

$$r_\text{max}^3 = \frac{3}{4\pi N_\text{V}}\,. \tag{27.31}$$

At this magnitude of the radius, the circular frequency of the orbiting electron has its smallest value, $\omega_{0,\text{min}}$. It is given by

$$\omega_{0,\text{min}}^2 = \frac{e^2 N_\text{V}}{3\varepsilon_0 m}\,. \tag{27.32}$$

equal to

$$\nu_{0,\text{min}}^2 = \frac{e^2 N_\text{V}}{12\pi^2\varepsilon_0 m}\,. \tag{27.33}$$

Below this limiting frequency, the electron is free. It can no longer be associated with a particular positive charge carrier or ion. We are then no longer dealing with oscillations within a single neutral molecule, but rather with oscillations of a whole set of electrons relative to all of the positive counter-ions, that is with *plasma oscillations*. $\nu_{0,\text{min}}$ *is the eigenfrequency of a (transversally) oscillating plasma.*C27.12 We insert it in place of ν_0 into Eq. (27.28) and obtain the dispersion formula of a plasma for the region of weakly-absorbed waves:

$$n^2 = 1 - \frac{e^2 N_\text{V}}{4\pi^2\varepsilon_0 m\nu^2} = 1 - 80.6\,\frac{\text{m}^3}{\text{s}^2}\cdot\frac{N_\text{V}}{\nu^2}\,. \tag{27.34}$$

C27.12. The derivation of the *plasma frequency*, i.e. the frequency of oscillations of a cloud of electrons within the lattice of positively-charged ions (as in a metal), can be found in Feynman's *Lectures on Physics*, Vol. II, Sect. 7.3 (available online, see Comments C6.1. and C7.1., and the solution to Exercise 26.1.), or also in F.S. Crawford, *Waves*, Berkeley Physics Course, Vol. 3 (McGraw Hill, New York 1968), p. 87. It is given by $\omega_\text{p}^2 = \dfrac{e^2 N_\text{V}}{\varepsilon_0 m}$.

27.19 The Total Reflection of Electromagnetic Waves by Free Electrons in the Atmosphere

In the case of *metallic bonding* (which occurs only in solids and in liquids), the strong interactions of closely-packed atoms (a high value of N_V!) leads to the formation of the best-known plasma: A cloud of freely mobile electrons within a lattice of positive ions. But Eq. (27.34) cannot be applied to metals due to their strong light absorption.

Electrons can however also be released without the interactions of closely-packed atoms, for example by the effects of ionizing radiations. Thus, especially through the action of ultraviolet light, electrons are set free in the upper layers of the atmosphere (the 'ionosphere'). Their number density at an altitude of 100 km is of the order of magnitude of $N_V = 10^{11}\,/\mathrm{m}^3$, and is thus very small compared to that in metals (e.g. $N_{V,\mathrm{Cu}} = 8.4 \cdot 10^{28}/\mathrm{m}^3$).

The *refractive index* produced by these free electrons can be calculated using Eq. (27.34). For an electron number density of $N_V = 10^{11}/\mathrm{m}^3$, equation (27.34) gives a refractive index which is very nearly equal to 1 for the frequency range of visible and infrared light (around 10^{15}-10^{12} Hz). However, in the region of *electric waves* ('radio' waves), the situation is different: At $\nu = 3 \cdot 10^6\,\mathrm{Hz}$ (corresponding to $\lambda = 100\,\mathrm{m}$), Eq. (27.34) gives $n = 0.32$, and thus a phase velocity of $9.4 \cdot 10^8$ m/s. For

$$\frac{\nu^2}{N_V} < 80.6\,\frac{\mathrm{m}^3}{\mathrm{s}^2} \quad \text{or} \quad N_V\lambda^2 > 1.12 \cdot 10^{15}\,\mathrm{m}^{-1}\,, \tag{27.35}$$

equation (27.34) even yields negative values for n^2, i.e. the refractive index becomes imaginary. Then even waves which are incident perpendicular to the layer are subject to total reflection[10]; no *travelling* wave can penetrate into the ionized layer. Making use of this total reflection, we can determine the number density of the electrons at various altitudes. A numerical example is given in Table 27.4. "Echoes" are rare for $\lambda < 30\,\mathrm{m}$. The necessary number density of the electrons would be $N_V > 1.8 \cdot 10^{12}/\mathrm{m}^3$; it seldom occurs, and then usually at altitudes of around 250 km.

The free electrons in the upper layers of the atmosphere (the 'ionosphere', e.g. the KENELLY-HEAVISIDE layer) are of great importance for communications by radio in the medium- and long-wave bands. They reflect the electromagnetic waves and guide them (along curved paths) to their distant targets. The lack of total reflection for $\lambda < 30\,\mathrm{m}$ makes it possible for short-wave electromagnetic waves that are emitted by the sun and more distant astronomical objects

[10] It follows from Eq. (25.15) for an imaginary refractive index n that the reflectivity is $R = (E_r/E_i)^2 = 1$ (numerator and denominator have the same magnitude).

Table 27.4 The total reflection of electromagnetic waves in the atmosphere

A signal of wavelength $\lambda =$	125 m	102 m
will, according to Eq. (27.35), be totally reflected at a number density of electrons $N_V =$	$0.7 \cdot 10^{11}/\text{m}^3$	$1.1 \cdot 10^{11}/\text{m}^3$
Its transit time t for outgoing and return paths is measured to be	$6.33 \cdot 10^{-4}$ s	$1 \cdot 10^{-3}$ s
The number density N_V leading to total reflection thus lies at the altitude $H_r = \frac{1}{2}tc =$	95 km	150 km

to reach the surface of the earth. They can be captured with large parabolic antennas (the concave 'mirrors' of radio telescopes). Radio astronomy has made many important contributions to our knowledge of the galaxy and the cosmos, e.g. the detection of the spiral structure of the Milky Way galaxy.

27.20 Extinction by Small Particles of Strongly-Absorbing Materials

In the cases described thus far, we have been able to treat separately the two contributions to extinction, i.e. scattering and absorption; the former in Chap. 26, and the latter in the present chapter. However, this separation is not always possible. This is the case for example with extinction by small particles which consist of strongly-absorbing materials.

Organic dyes and metals exhibit *strong absorption* even in the visible spectral region. When finely divided, they have quite different extinction spectra than as bulk materials. A long-known example is *ruby glass*. It contains very finely-divided gold particles; however, it does not allow green light to pass through, like very thin gold leaf, but instead red light (Fig. 27.20). The diameter of the individual gold particles is below the resolving power of light microscopes, but each particle produces a colored diffraction disk under dark-field illumination in the viewing field of a microscope. Light is thus scattered from each particle[11]. The proportions of scattering and absorption are found from experience to depend strongly on the size of the particles: Very small particles scatter only weakly; they attenuate the light mainly by absorption.

For a quantitative investigation, a solid solution of sodium in an NaCl crystal is suitable. A hot NaCl crystal in Na vapor takes up excess Na atoms. The mechanism of this process is known: A small fraction of the negative chlorine *ions* in the lattice is displaced (and replaced) by thermally-diffusing *electrons*. The resulting absorption centers are

[11] This detection of single particles is called "ultramicroscopy".

Figure 27.20 The extinction spectrum of gold ruby glass

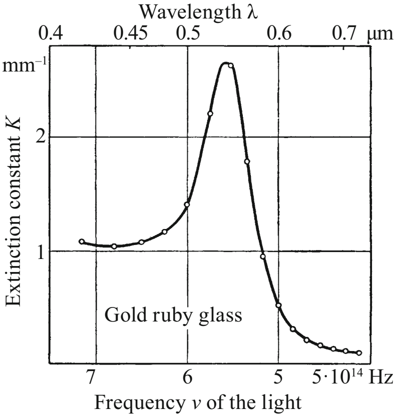
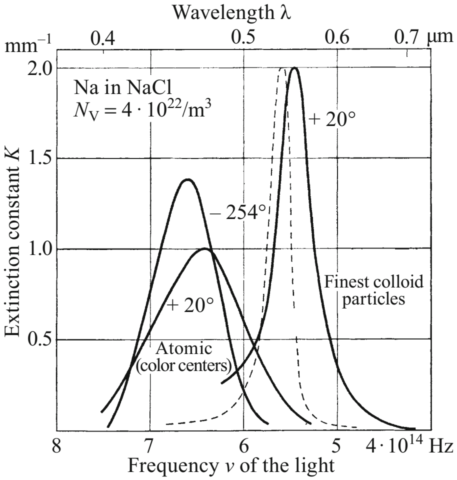


Figure 27.21 The extinction spectra of atomic and colloidally-dissolved metal particles (Na in an NaCl crystal). The dashed curve for the finest colloid, which shows no scattering, was calculated using Eq. (27.37).



called *color centers* (see Sect. 27.14). In equilibrium, the number density N_V of Na atoms in the crystal is about the same as in the vapor; at 500 °C, for example, it is $N_V = 5 \cdot 10^{22}/\text{m}^3$. In thermal equilibrium at room temperature, $N_V = 3 \cdot 10^{11}/\text{m}^3$ would be found in the crystal. Such small number densities cannot be detected even by absorption-spectrum analysis (Sect. 27.15). Therefore, one has to "quench" the crystal and "freeze in" the number density obtained at a high temperature, so that it remains stable at room temperature. Figure 27.21, left, shows the extinction spectrum of such a "frozen in" atomic solid solution of Na (color centers) in an NaCl crystal, at two temperatures. The extinction is due here entirely to *absorption*. No trace of *scattering* is observable.

At room temperature, the frozen-in number density in an NaCl-crystal can be maintained for years. At 300 °C, on the other hand, the diffusion rate has become measurably large. As a result, the crystal lattice can precipitate a portion of the excess sodium, allowing it to condense into fine colloidal particles. This lowers the color-center band; at the same time, a new extinction band with a maximum at 0.550 µm appears (at the right in the figure). The extinction in this band is also practically due only to absorption and not to scattering. Its position remains nearly unchanged with changing temperature, in contrast to that of the color-center band. Upon longer heating, the particles increase in size, and their extinction bands are shifted and extended to the region of longer wavelengths. Only then does scattering begin to play a role in the crystal, initially weakly and then more strongly.

The maximum of the new band (measured at room temperature) always lies at least 0.08 µm to the long-wavelength side of the maximum of the color-center band. There is thus no gradual transition of the color-center band by a continuous shift into the new band. As a result, we must ascribe the new band to the *smallest* colloidal particles that remain stable.

For the *atomically*-dissolved metals in alkali halide crystals (color centers), the shape of the band can be represented by the model of damped resonators (Fig. 27.18). The position of the band is determined by the lattice constant a of the crystals (cf. the footnote at the end of Chap. 22). The frequency of its maximum at 20°C is given by an empirical relation[12]:

$$\nu_{\max} \cdot a^2 = 2.02 \cdot 10^{-4}\,\mathrm{m^2/s}\,. \qquad (27.36)$$

For the *colloidally*-dissolved metals, in contrast, the shape and position of the band are determined by the *optical constants of the metals* (indeed, by the values of n and k measured on *bulk* samples), and not by damped resonators. With these constants, we can calculate the absorption constant K for the smallest colloidal particles (diameter $\ll \lambda$) at various wavelengths. The following equation is used for the calculation:

$$K = 36\pi N_\mathrm{V} V \frac{1}{\lambda} \cdot \frac{\dfrac{nk}{n_\mathrm{u}}}{\left[\left(\dfrac{n}{n_\mathrm{u}}\right)^2 + \left(\dfrac{k}{n_\mathrm{u}}\right)^2\right]^2 + 4\left[\left(\dfrac{n}{n_\mathrm{u}}\right)^2 - \left(\dfrac{k}{n_\mathrm{u}}\right)^2 + 1\right]} \qquad (27.37)$$

(Here, n_u is the refractive index of the *solvent*, λ the wavelength in air, N_V the number density of the particles, and V is the volume of the individual particles. For a derivation, see earlier editions of this book[C27.13]).

C27.13. See the 9th edition (1954), or the 10th edition (1958) of "*Optik und Atomphysik*", p. 206.

For our 'standard example', the smallest Na colloid in an NaCl crystal, the optical constants of sodium are collected in Fig. 27.22 (top). n_u, the refractive index of the surroundings, that is of the NaCl crystal, is practically constant at 1.55 (Fig. 27.1, bottom). There are no

[12] E. Mollwo (Dr. rer. nat. Göttingen, 1933), *Zeitschrift für Physik* **85**, 56 (1933).

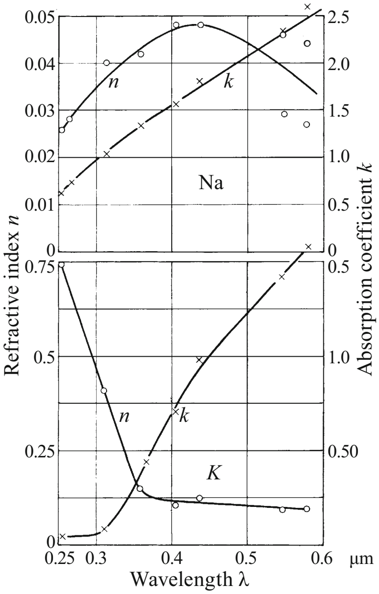


Figure 27.22 The optical constants of sodium and potassium. At $\lambda < 0.31\,\mu\text{m}$, potassium has a region of weak extinction, i.e. $k < 0.1$ (Sect. 25.5); but the extinction constant K is still about $2 \cdot 10^3\,\text{mm}^{-1}$. For rubidium and cesium, the curves are similar to those for potassium. Therefore, for sodium also, we can expect a steep rise in the refractive index n at $\lambda < 0.25\,\mu\text{m}$.

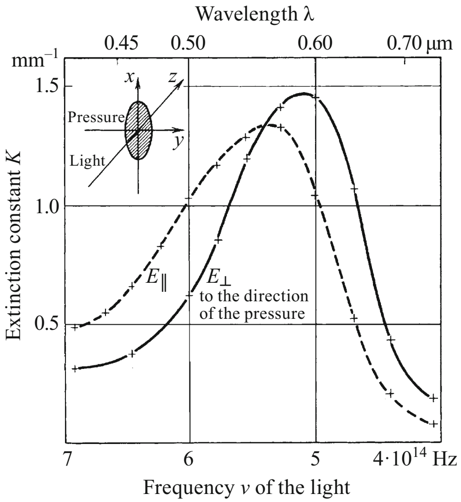
reliable data for N_V and V; thus, we just compute the right-hand product in Eq. (27.37) for various values of λ. This leads us to the dashed curve in Fig. 27.21. Its maximum value is fitted to the observed value by choosing the constant in Eq. (27.37). n and k are hardly temperature dependent, and the same holds for the computed function. Result: The calculation can correctly reproduce the two essential features of light extinction by very fine colloidal metal particles, namely the small widths of the bands and their lack of temperature dependence. Furthermore, the respective frequency of each band maximum nearly coincides with the measured value[13] in each case. The remaining discrepancies are not alarming. They could be eliminated by small changes in the interpolation curves for n and k.

27.21 Extinction by Large Metal Colloids. Artificial Dichroism, Artificial Birefringence

For very fine metal and dye colloids, no secondary radiation can be observed, only absorption. Only when the colloids consist of larger particles (diameters or circumferences comparable to λ) do we observe secondary radiation and scattering in addition to absorption. In this case, the individual parts of the colloid particles are not all

[13] The extinction curves of the fine colloids in Fig. 27.21 are not "optical resonance curves"; their shapes are instead determined by the curves of the optical constants of the material of the particles.

Figure 27.23
Pressure-induced dichroism in NaCl containing colloidal Na particles



excited with the same phase by the primary waves. As a result, interference occurs, giving preferred directions to the secondary waves; in particular in the direction of the primary waves, so that forward scattering predominates (compare Fig. 26.10). We can thus no longer assume a model employing the simple electric polarization of small spheres for a quantitative description of this process. Instead, we must use a similar description as in the calculation of the harmonics of antennas. In this calculation, the essential quantity is the shape of the particles; but precisely this is usually unknown for large colloidal particles.

We of course cannot explore these complicated matters in detail here; rather, we must content ourselves with a qualitative treatment of *artificial dichroism* (Sect. 24.3). As an illustration, we use a large-diameter Na colloid in an NaCl crystal. The crystal appears violet in transmitted light, and yellow-brown in reflected light. Its broad extinction band has a maximum around 0.59 μm, independent of the orientation of the plane of oscillation of polarized light.

Now we subject the crystal to pressure parallel to one of its cubic edges. The result is that it becomes *dichroitic*, i.e. it now exhibits two overlapping extinction bands in polarized light (Fig. 27.23). Explanation: The pressure causes the colloidal particles to take on an elongated shape (inset in Fig. 27.23). In the case of $E_\perp$, the amplitude oscillates parallel to the longer particle axis x, while in the case of $E_\parallel$, it oscillates parallel to the shortest axis y. For $E_\perp$, the long diameter of the particles preferentially determines the wavelength, while for $E_\parallel$, the shortest diameter is determining.

All birefringent substances are dichroitic; this follows inevitably from the general relation between dispersion and absorption. This connection is represented schematically in Fig. 27.24. The full curves refer to one of the two polarized oscillations, and the dashed curves

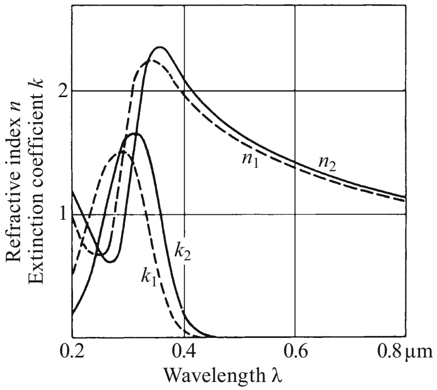


Figure 27.24 A schematic sketch illustrating the dichroism of all birefringent substances

to the other, which is perpendicular to the first. With transparent substances (calcite, mica, quartz), both absorption spectra end already in the ultraviolet spectral region, before the beginning of the visible.

The fabrication of very thin birefringent crystal platelets is rather difficult. Therefore, the relevant absorption bands for birefringence have been measured only in a few cases. With artificial dichroism, the concentration of the light-absorbing particles is low, so that one need not bother with very thin crystal samples. Its disadvantage is however that now, the birefringence produced by the particles is quite small, and furthermore it is seen on the background of the birefringence of the strained solid solvent (Sect. 24.9). Therefore, we cannot detect the birefringence from parallel-oriented, elongated particles with certainty by simple means. That can however be accomplished in other cases.

A parallel orientation of small particles can also be achieved in numerous ways even for large particle-number densities; among others, by using electric fields or laminar flows in fluids. For example, we can place a few drops of a suspension of vanadium pentoxide (V_2O_5) in water between two glass plates and slide the plates relative to each other by several millimeters. Immediately, the liquid layer becomes birefringent ("flow birefringence"). It acts in the setup of Fig. 24.16 just like a crystal plate G. Still more impressive is the demonstration experiment described in Fig. 27.25.

Artificial birefringence can also be produced using polar molecules, as well as nonpolar molecules which however can be strongly deformed by electric fields. The best-known examples are nitrobenzene and carbon disulfide. The crystal platelet G in Fig. 24.16 is replaced by a parallel-plate condenser filled with one of these liquids, the field direction is adjusted to be perpendicular to the light beam, and field strengths E on the order of 10^4 V/cm are employed. This form of artificial birefringence was discovered by J. Kerr in 1875. Experimentally, at a wavelength λ, one finds a difference of the refractive

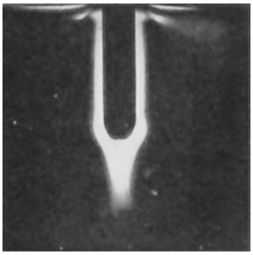


Figure 27.25 A demonstration experiment showing flow birefringence (photographic positive). A glass cuvette about 4 cm deep and 1 cm thick is filled with a suspension of V_2O_5 in water and observed between crossed NICOL prisms (Sect. 24.3; see Fig. 24.16). When a glass rod is dipped into the liquid, the layers which flow around it glow with a bright red color. Similarly, when the liquid is stirred, the turbulence becomes visible; and likewise, in a tube with a laminar flow, the boundary layer at rest on the walls of the tube can be seen.

indices for the extraordinary and the ordinary light beams equal to

$$n_{\mathrm{eo}} - n_{\mathrm{o}} = B \cdot \lambda \cdot E^2 \,. \tag{27.38}$$

In this equation, the "electric KERR constant" B is given by

$$B = \frac{n_{\mathrm{eo}} - n_{\mathrm{o}}}{\lambda \cdot E^2} = \frac{\Delta}{\lambda} \cdot \frac{1}{l} \cdot \frac{1}{E^2} \,, \tag{27.39}$$

when the light traverses a path l in the electric field and thereby experiences an optical path length difference of $\Delta = (n_{\mathrm{eo}} - n_{\mathrm{o}})\, l$.

Explanation: The molecules which exhibit a KERR effect have an asymmetric structure. They have a preferred direction of polarizability. The dipole moments produced by the field are proportional to the field strength E. Furthermore, the polarized molecules are progressively oriented as the field is increased, gradually overcoming their random thermal motions. Therefore, the birefringence increases proportionally to E^2.

Numerical example: For very pure nitrobenzene, $B = \dfrac{4.3 \cdot 10^{-10}}{\mathrm{cm}(\mathrm{V/cm})^2}$. Then, for $l = 1\,\mathrm{cm}$ and $E = 10^4\,\mathrm{V/cm}$: $\dfrac{\Delta}{\lambda} = B \cdot l \cdot E^2 = \dfrac{4.3 \cdot 10^{-10} \cdot 1\,\mathrm{cm} \cdot 10^8 (\mathrm{V/cm})^2}{\mathrm{cm}(\mathrm{V/cm})^2} = 4.3 \cdot 10^{-2}$ or $\Delta \approx 0.04\,\lambda$.

The KERR effect is employed technically for constructing control devices ('switches') for light. The radiant power transmitted by the device increases initially as $\sim E^4$.

Exercises

27.1 Derive the relationship between the halfwidth H and the logarithmic decrement Λ which is used in Eq. (27.16) for a linear mass-and-spring pendulum (Fig. 4.13 in Vol. 1). Consider the case of weak damping ($\Lambda < 1$), and thus a narrow resonance curve. In this case, the amplitude l_0 in Eq. (26.1) can be approximated as $l_0 \approx F_0 / \left[4\pi^2 m \sqrt{(2\nu_0^2)(\nu_0 - \nu)^2 + (\Lambda/\pi)^2 \nu_0^4} \right]$, as can be easily shown (for a narrow resonance curve, $\nu \approx \nu_0$). Also, take into account that the energy of the oscillator is proportional to the square of the amplitude.
(Sect. 27.13)

27.2 A linear mass-and-spring pendulum with a mass m vibrates at the circular frequency ω. Its amplitude is l_0. Find the mean value of its kinetic energy $\overline{W}_{\text{kin}}$ (Eq. (27.17)).
(Sect. 27.13)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_27) contains supplementary material, which is available to authorized users.

Thermal Radiation 28

28.1 Preliminary Remark

Among the various possibilities for excitation of molecules and atoms, *thermal excitation* has played a special role since ancient times. Therefore, thermally-excited radiation (or "temperature radiation") has been extensively investigated. The crowning achievement of this work was the formulation and derivation of PLANCK's thermal radiation law in 1900, and with it the discovery of the physical constant h, PLANCK's constant or PLANCK's quantum of action.

28.2 Basic Empirical Results

The fundamental experimental results can be summarized briefly:

1. *All objects radiate energy towards each other. Warmer objects are thus cooled, while cooler objects are warmed.* In order to demonstrate this transfer of heat via radiation, we must first avoid thermal *conductivity*. It is expedient to use two concave mirrors facing each other at a distance of several meters. At the focal point of one mirror, we place a radiometer (a radiation thermopile). At the focus of the other, we can first hold a warm finger, then a beaker filled with ice water. In the first case, the radiometer will indicate warming, and in the second, it will indicate cooling (jokingly known as 'cool radiation').

2. *The radiant intensity increases strongly with increasing temperature*. This can be demonstrated using an electric cooking pot with a thermometer, which is placed as a "radiating emitter" at a distance of around 0.5 m from a radiometer that serves as "receiver".

3. *As the temperature increases, the* distribution *of the radiant intensity in the thermal spectrum also changes*. With a slowly heated metal wire, we can demonstrate the sequence: First invisible infrared radiation, perceptible only as heat; then glowing red, glowing yellow, finally glowing 'white hot'.

4. *At a given temperature, an object which absorbs light emits more thermal radiation than a transparent or a strongly reflecting object*. To demonstrate this, we heat different but equal-sized objects beside each other in similar colorless BUNSEN-burner flames and observe the light that they emit: A rod made of clear glass absorbs practically no visible light and emits only weakly. A rod made of colored

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_28

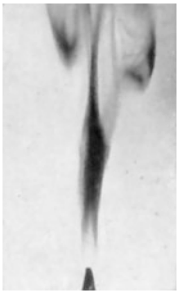


Figure 28.1 A brightly glowing, strongly turbulent flame from natural gas containing benzene vapor casts a deep dark shadow when placed in front of the condenser of a projector

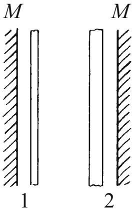
glass absorbs part of the visible spectrum and emits strongly. A clear glass tube, filled with finely-powdered colored glass, *scatters* incident light. Only a small fraction of the light can penetrate into the interior of the tube and be absorbed there. The powder thus *absorbs* less than the solid colored rod, and therefore, it also *emits* less than the rod.

Another example: A brightly glowing flame of "carbureted" natural gas which contains benzene vapor is placed in front of the condenser of a projector: A deep, dark image of the flame then appears on the screen (Fig. 28.1). The innumerable fine carbon particles (soot) which float in the gases of the flame absorb a noticeable portion of the light from the projector lamp. Now, by increasing the air flow, we convert the flame to a colorless BUNSEN-burner flame in the well-known manner; i.e. all the carbon is burned and no soot is formed. Then we no longer see an image of the flame on the screen, it no longer absorbs visible light. At the same time, its emission also vanishes. A flame which can absorb no visible light also cannot emit visible light. A candle flame likewise produces a dark image in a projector. In general, thermal emission of incandescent light by a flame is based on the presence of solid particles which *absorb visible light*, namely soot particles.

28.3 KIRCHHOFF's Law

Quantitatively, the experimental facts listed above can be described by KIRCHHOFF's law. We explain its content by considering a thought experiment. In Fig. 28.2, the objects 1 and 2 represent small sections of two large, flat objects. They are made of two different, arbitrary materials. Each of them radiates thermal energy towards the other, and in equilibrium they are at the same temperature. The radiated power which is emitted from their back surfaces is reflected completely and without losses by the two ideal mirrors M. Thus we need consider only the radiation *between* the two objects. In a steady state, object 1 must radiate just the same radiant power to object 2 as object 2 receives. Object 1 radiates its own radiant power $\dot{W}_1$ towards 2, and in addition it reflects the non-absorbed fraction

Figure 28.2 Illustrating KIRCHHOFF's law



$(1 - A_1)$ of the power $\dot{W}_2$ that it receives from 2. Here, A_1 is the *absorbance* of the object for non-monochromatic radiation, defined by the equation

$$\text{Absorbance } A = \frac{\text{Absorbed radiant power}}{\text{Incident radiant power}} . \qquad (28.1)$$

The corresponding conclusions hold conversely for the radiation emitted by 2 and transported to 1. Therefore, in equilibrium, we have

$$\dot{W}_1 + (1 - A_1)\dot{W}_2 = \dot{W}_2 + (1 - A_2)\dot{W}_1 ,$$

that is

$$\frac{\dot{W}_1}{A_1} = \frac{\dot{W}_2}{A_2} ,$$

and, in general,

$$\frac{(L_e)_1}{A_1} = \frac{(L_e)_2}{A_2} , \qquad (28.2)$$

where L_e denotes the radiance (Sect. 19.2). This relation holds for any two arbitrary bodies. Therefore, the quantity L_e/A must be independent of all *materials properties*. It can depend only on *other* quantities, such as the temperature or the wavelength of the radiation. This statement is KIRCHHOFF's law.

Continuing our thought experiment, we suppose that between the objects 1 and 2 there is now an absorption-free interference filter (Sect. 20.13) which allows only a very narrow interval of wavelengths $\Delta\lambda$ to pass through. The radiance is given by

$$L_e = \int_0^\infty \frac{\partial L_e}{\partial \lambda} \cdot \mathrm{d}\lambda$$

($\partial L_e / \partial \lambda$ is the *spectral* radiance),

so that for the selected wavelength interval $[\lambda \to \lambda + \Delta\lambda]$ (which we denote simply by its central "wavelength λ"), we have

$$L_{e,\lambda} = \int_\lambda^{\lambda+\Delta\lambda} \frac{\partial L_e}{\partial \lambda} \cdot \mathrm{d}\lambda ;$$

and, instead of Eq. (28.2), we find

$$\frac{(L_{e,\lambda})_1}{(A_\lambda)_1} = \frac{(L_{e,\lambda})_2}{(A_\lambda)_2}\,. \tag{28.3}$$

An object 1 with the absorbance $(A_\lambda)_1 = 1$ thus absorbs *all the incident radiation* of wavelength λ; it is termed *black* for λ. Then it follows from Eq. (28.3) that

$$(L_{e,\lambda})_2 = (L_{e,\lambda})_1 \cdot (A_\lambda)_2\,. \tag{28.4}$$

In words: For thermally-excited, monochromatic radiation, the radiance $(L_{e,\lambda})_2$ of any arbitrary body is equal to the radiance $(L_{e,\lambda})_1$ of a body which is "*black*" for the wavelength λ, multiplied by the absorbance $(A_\lambda)_2$ of the body which is not black at λ.

28.4 The Black Body, and the Laws of Black-Body Radiation

Zero light reflectivity, i.e. an absorbance of $A = 1$, can be realized in the form of a small opening in the wall of a box which is otherwise opaque to light. Such an opening appears blacker than a soot-coated plate held next to it. It absorbs all of the incident light, via multiple, mostly diffuse reflections. Following a suggestion by G. Kirchhoff (1859), such black bodies were heated to high, uniformly distributed temperatures and their openings were employed as emission sources ("black-body radiators"). The incandescent light which emerges from the *opening* is called *black-body radiation*.

As a demonstration experiment, we electrically heat a roughly 15 cm long platinum tube of ca. 2 cm diameter in air until it begins to glow. A weakly reflecting cross is drawn on the wall of the tube using iron oxide paint. Near it, the tube wall has a small opening to the interior. The polished, highly reflective platinum tube wall glows the least; the weakly-reflecting cross glows more strongly, but the completely non-reflecting "black" opening glows the most brightly.

Larger black bodies can be constructed of fireproof ceramic materials. In general, a long tube with several cross-sectional baffles is sufficient. Its outer wall is covered with an insulating material in order to conserve heating energy. For the purposes of measurements at high temperatures, tungsten is a suitable material. It is mounted and heated just like the tungsten filaments in an incandescent lamp; that is, no external insulation is required.

An essential feature of every usable black body is that the temperature in its interior must be uniformly distributed and constant. If this condition is fulfilled, then, looking in through an opening, we can see no details or internal structure. Examples are the furnaces used to

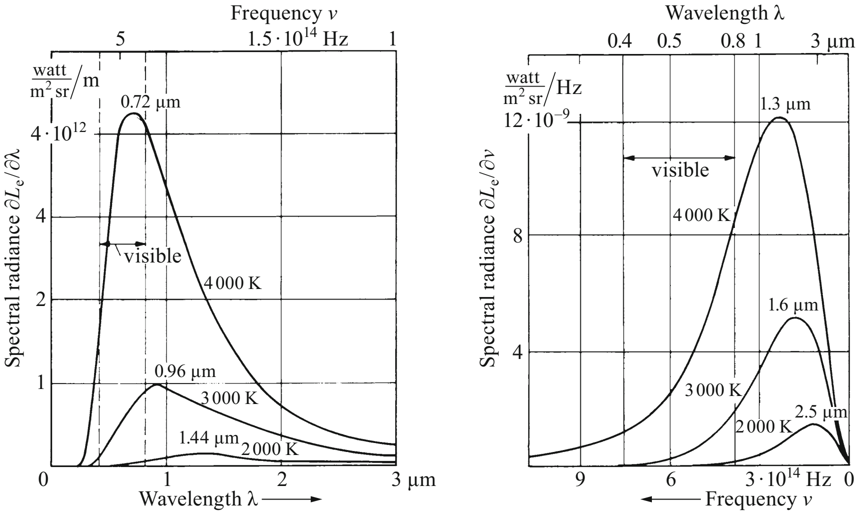


Figure 28.3 The spectral radiance distribution in the spectrum of a black body. At the left, it is referred to equal wavelength intervals, and at the right, to equal frequency intervals. These curves, as well as Eqns. (28.5) and (28.6), hold for *unpolarized* radiation. For the quantitative experimental investigation of the spectral radiance distribution, one uses the arrangement sketched in Fig. 19.1 (usually for the special case of $\vartheta = 0$, that is perpendicular emission). One measures the radiant power $\mathrm{d}\dot{W}_\lambda$ or $\mathrm{d}\dot{W}_\nu$ which is emitted into the selected spectral interval and into the solid angle $\mathrm{d}\Omega$. From the defining equation for the radiance L_e, it then follows that

$$\mathrm{d}\dot{W}_\lambda = \frac{\partial L_\mathrm{e}}{\partial\lambda}\cdot\mathrm{d}\lambda\cdot\mathrm{d}A_\mathrm{P}\cdot\mathrm{d}\Omega \quad \text{or} \quad \mathrm{d}\dot{W}_\nu = \frac{\partial L_\mathrm{e}}{\partial\nu}\cdot\mathrm{d}\nu\cdot\mathrm{d}A_\mathrm{P}\cdot\mathrm{d}\Omega$$

($\mathrm{d}A_\mathrm{P}$ is the projected emitting area as in Fig. 19.3. For perpendicular emission, $\mathrm{d}A_\mathrm{P} = \mathrm{d}A$).

melt glass (glass kilns) in a glass factory, or the coke ovens in a coking plant. Every surface element in the interior of the black body, independently of its material or structure, emits at exactly the same radiance: Surface elements with a high absorbance (Eq. (28.1)) also *emit* a large amount and *reflect* very little of the radiation from all the other surface elements. For surface elements with a low absorbance, the converse holds: they themselves emit less strongly, but they reflect more of the incident radiation from the other surface elements.

The distribution of the radiance over the various spectral intervals has been investigated quite thoroughly for "black-body" radiation, that is, for the incandescent light emerging from the opening of a black body. In particular, its dependence on the temperature has been very carefully determined. The results are collected in Fig. 28.3. The spectral radiance is plotted in the left graph against equal wavelength intervals, and in the right graph against equal frequency intervals. In the search for a formal description of these empirical results, a number of competent physicists invested considerable efforts. The final success was attained at the end of 1900 by MAX PLANCK, with his famous radiation formula[1]:

$$\frac{\partial L_\mathrm{e}}{\partial \lambda} = \frac{C_1}{\lambda^5} \cdot \frac{1}{e^{\frac{C_2}{\lambda T}} - 1} \tag{28.5}$$

or

$$\frac{\partial L_\mathrm{e}}{\partial \nu} = C_3 \cdot \nu^3 \cdot \frac{1}{e^{\frac{C_4 \nu}{T}} - 1} \,. \tag{28.6}$$

$C_1 \,\ldots\, C_4$ are empirical coefficients with the following values:

$$C_1 = 1.191 \cdot 10^{-16}\,\mathrm{W\,m^2}\,, \qquad C_2 = 1.439 \cdot 10^{-2}\,\mathrm{m\,K}\,,$$

$$C_3 = 1.47 \cdot 10^{-50}\,\frac{\mathrm{W\,s^4}}{\mathrm{m^2}}\,, \qquad C_4 = 4.78\,10^{-11}\,\mathrm{s\,K}\,.$$

PLANCK wanted to refer these coefficients back to universal physical constants. In this process, he made one of the greatest physical discoveries: He found the new universal physical constant h. PLANCK was the first to introduce the energy equation $E = h \cdot \nu$, and with it, he opened the door to the world of atomic-scale processes (now called quantum physics).

Today, there are several derivations of PLANCK's formula. We refer to the details given in all modern textbooks on theoretical physics. However, independently of the derivation, the relationship between the empirical constants in the radiation formula and the universal physical constants remains. We find:

$$C_1 = 2hc^2\,, \quad C_2 = \frac{hc}{k}\,, \quad C_3 = \frac{2h}{c^2}\,, \quad C_4 = \frac{h}{k}$$

(h is PLANCK's constant $= 6.626 \cdot 10^{-34}$ W s^2, k is BOLTZMANN's constant $= 1.38 \cdot 10^{-23}$ W s/K, and c is the vacuum velocity of light $= 3 \cdot 10^8$ m/s).

PLANCK's radiation formula contains two important laws as special cases; these had both been discovered previously:

1. The STEFAN-BOLTZMANN *law*: The total power emitted outwards from a surface area $A^{\mathrm{C28.1}}$ of a black body increases proportionally to the 4th power of its temperature T, that is

C28.1. In this chapter, the letter A is used both for the absorbance of a body and also for a surface area. This should however be clear from context, so that the danger of confusing the two quantities is minimal.

$$\dot{W} = \sigma \cdot A \cdot T^4 \tag{28.7}$$

$$\left(\sigma = \frac{2\pi^5 k^4}{15 c^2 h^3} = 5.67 \cdot 10^{-8}\,\frac{\mathrm{W}}{\mathrm{m^2\,K^4}}\right)\,.$$

[1] In the visible spectral range, i.e. for $\lambda < 0.8\,\mu\mathrm{m}$, and up to $T = 3000\,\mathrm{K}$, the term -1 in the denominator can be left off; the resulting error is less than 0.1 % (this then gives the radiation formula of W. WIEN).

The sun emits radiation to a good approximation as a black body. At its surface, the emitted power per surface area (Sect. 19.3) is:

$$\frac{\dot{W}}{A} = \pi L_{\mathrm{e}} = 6.3 \cdot 10^7 \, \frac{\mathrm{W}}{\mathrm{m}^2} \, .$$

According to Eq. (28.7), this corresponds to a temperature of 5700 K. See Fig. 28.6.

In practical applications of this equation, it is often desirable to determine the net power which is lost by a body through radiation. Then, in addition to the power *emitted* by the body, we must also consider the power with which the body is *irradiated* from its surroundings. This reduces the net power loss by the body. The result is

$$\dot{W} = \sigma \cdot A \left(T^4 - T_{\mathrm{s}}^4\right) \tag{28.8}$$

(T_{s} is the temperature of the surroundings, e.g. of the receiver).

2. The *displacement law* of W. WIEN: The wavelength $\lambda_{\max}$ at which the spectral radiance has its maximum value is inversely proportional to the temperature T. We have:

$$\lambda_{\max} \cdot T = \frac{hc}{4.97 \cdot k} = 2.88 \cdot 10^{-3} \, \mathrm{m\,K} \, . \tag{28.9}$$

In the solar spectrum, we can observe that the maximum value of the spectral radiance is at the wavelength $\lambda = 0.48\,\mu\mathrm{m}$. For a black body, this corresponds to a temperature of 6000 K (see Fig. 28.6).

28.5 Selective Thermal Radiation

The absorbance A of a black body is 1 for all wavelengths. For all other bodies, A depends on the wavelength and is furthermore always less than 1. For this reason, at a given temperature and wavelength, instead of the radiance $L_{e,\lambda}$ of a black body, only the fraction $A \cdot L_{e,\lambda}$ is emitted by a non-black object. A is especially small for the limiting cases of *very strong* or *very weak* absorption (Sect. 25.5). With strong absorption ($w < \lambda$ as in metals, $w =$ average penetration depth of the radiation), the radiation cannot penetrate into the object; often, more than 90 % of the incident power is reflected. For "weak absorption" ($w > \lambda$), only a small fraction of the radiation is prevented from penetrating the object by reflection, and therefore, the major portion of the incident radiation can be absorbed. This however takes place at great depths, too thick for many technical applications. There is in addition a further complication: The optical constants also change with temperature.

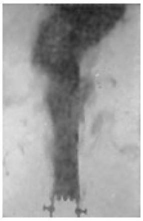


Figure 28.4 ZnO smoke, glowing blue-green. In the light of an arc lamp, the smoke casts a deep black shadow, just like the soot particles in a glowing gas or candle flame (Fig. 28.1).

This dependence is well understood in only a few cases and for limited spectral ranges, e.g. for metals in the infrared. There, the reflectivity R is determined only by the electrical conductivity of the metal (Sect. 27.17), and its temperature dependence is well known. In general, for non-black bodies, the dependence of the radiance $L_{e,\lambda}$ on λ can only be determined experimentally, and only approximately. Very few materials can survive large variations of their temperatures without permanent changes. Nearly always, the internal and surface structures depend strongly on the thermal history of the object. A microcrystalline texture is converted into a rough mosaic of strongly reflecting single crystals, etc.

In the visible spectral range, selective thermal emission can be readily demonstrated experimentally: A small quartz-glass plate is half covered with a thin layer of ZnO, and the other half is covered with a Pt film. On heating over a Bunsen burner, the platinum begins to glow red, but the ZnO glows blue-green. The reason is that hot ZnO crystals absorb only the short-wave portion of the visible spectrum, with a very steeply rising absorption curve; as a result, they can emit only this part of the spectrum thermally. To make this demonstration visible to a large audience, we heat a zinc-coated iron wire electrically (Fig. 28.4): The zinc vaporizes, the vapor becomes oxidized, and the hot ZnO smoke glows like a blue-green torch, visible from a considerable distance.

28.6 Thermal Light Sources

Thermal light sources make exclusive use of the radiation from *solid* bodies. These are heated either by chemical processes (in a flame), or electrically by Joule heating. There are in principle two ways of shifting a large fraction of the emitted radiant power into the visible spectral range: *High temperatures* (Fig. 28.3), and the use of *selective emitters*. Their absorbance must be as close as possible to that of a black body in the visible range, and as small as possible in other spectral regions, especially in the infrared.

The flames which have been used as light sources since antiquity produce a typical *incandescent gaslight*: Solid objects (kindling, torches) or liquid fuels, soaked up by a wick, are converted by the heat of combustion into gaseous hydrocarbons. These are not completely burned. *Solid* carbon is formed as very fine particles, called

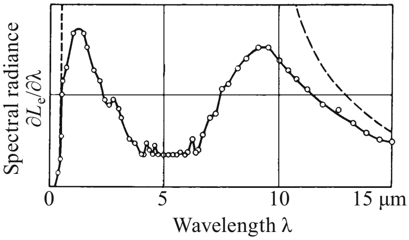


Figure 28.5 The spectral radiance. The solid curve is for an AUER mantle, while the dashed curve applies to a black body at the same temperature

soot (Fig. 28.1). These solid, very hot carbon particles produce the radiation.

Only at the beginning of the 19th century was the production of gas for combustion separated from the locality of the consumer. The gas was produced centrally from solid or liquid fuels and distributed to the consumers through pipes. The last decade of the 19th century then saw the second leap of progress since PROMETHEUS:$^{C28.2}$ The carbon particles in the flame, which radiate nearly as "black bodies" in the infrared as well as the visible range, were replaced by a *mantle* which radiates *selectively*. It is heated by a colorless BUNSEN-burner flame and emits predominantly visible radiation.

"The last decade of the 19th century then saw the second leap of progress since PROMETHEUS".

C28.2. According to ancient Greek mythology, PROMETHEUS brought fire back to mankind using a torch ignited secretly on the Sun Chariot; it had previously been taken away by ZEUS.

> The mantle, a small bag of silk mesh, is soaked with a suspension of very selectively absorbing cerium oxide (ca. 1 %) in a thin and therefore not strongly absorbing layer of thorium oxide. Fig. 28.5 shows the spectral radiance $\partial L_e/\partial\lambda$ from a commercial AUER mantle (C. AUER, 1885; $T \approx 1800\,\mathrm{K}$), and above it, dashed, the form of the $\partial L_e/\partial\lambda$ curve of an ideal black body at the same temperature. In the blue spectral range, the curves coincide; there, the absorbance A of the AUER mantle is nearly equal to 1. Thus, the mantle radiates nearly like a black body in this range. Between $\lambda = 1$ and $7\,\mu\mathrm{m}$, however, the absorbance A of the mantle is small, and therefore only a small portion of the radiance is emitted in this spectral region, which is useless for illumination purposes. For $\lambda > 9\,\mu\mathrm{m}$, the radiance again approaches that of a black body.

When JOULE heating is to be used to raise the temperature, today we make use of *metal* wires with a high melting point. Metals have a high reflectivity R in the infrared region (Fig. 27.8) and thus a small absorbance $A = 1 - R$ there. As a result, their thermally-emitted radiation also predominates at shorter wavelengths. Temperatures of the order of 6000 K would be desirable (Fig. 28.6). But even tungsten, with its melting point of $T_m = 3700\,\mathrm{K}$, can tolerate at most temperatures of only about 2700 K over longer periods of time due to increased evaporation of the metal. This is the usual operating temperature of gas-filled tungsten-filament lamps with a double-wound filament (Fig. 28.7). The windings of the filament radiate nearly as black bodies in the visible. Their lifetime is longer than 1000 hrs. For tungsten lamps which produce a particularly high radiance L_e, the temperature is increased up to around 3400 K. But the operating lifetime of the filament is then only 1 to 2 hrs. In both types of lamps, evaporation of the metal filament must be reduced by using an unre-

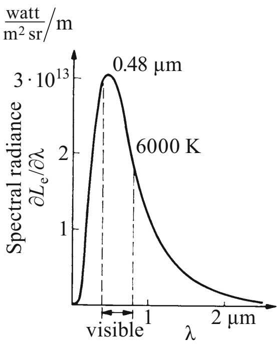


Figure 28.6 The distribution of the spectral radiance of a black body at 6000 K, the temperature of the sun's surface

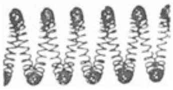


Figure 28.7 A double-wound lamp filament

active gas atmosphere (Ar, Kr) or a halogen-containing gas ("halogen lamps").

Modern developments in illumination technology replace thermal excitation by electrical excitation. The well-known and widespread "fluorescent tubes" in use today are filled with mercury vapor (or in some cases noble gases) and are a further development of the very first form of electrical illumination (Francis Hauksbee, 1705[C28.3]). In contrast to thermally-excited sources, these fluorescent tubes produce a large amount of ultraviolet radiation, which is useless to the human eye for illumination. It is converted to visible radiation by coating the inner walls of the glass tubes with a chemical compound that absorbs the ultraviolet light and emits visible fluorescence light. For almost colorless fluorescent tubes, the light emission efficiency is (80–100) lm/W[C28.4], and thus around five times greater than that of gas-filled tungsten-filament lamps. We should also mention here the electroluminescence from solid-state *light-emitting diodes* (LEDs). Their performance is improving rapidly; their light emission efficiency is currently over 100 lm/W.

C28.3. Francis Hauksbee, F.R.S. (died around 1713). He produced light by shaking an evacuated glass tube partially filled with mercury. (*Philosophical Transactions*, No. 303, Vol. 24 (1705); see also J.L. Heilbron, "*Electricity in the 17th and 18th centuries: A study of early modern physics*", Univ. of California Press, Berkeley (1979), Chap. VIII.)

C28.4. The unit 'lumen' (lm) is defined in Table 29.1.

28.7 Optical Thermometry. The Black-Body Temperature and the Color Temperature

Black-body radiation and its laws find an important application in the measurement of temperatures above 600 °C. Above 2600 °C, optical thermometry is in fact the only practicable method of temperature measurement[2].

[2] Gas thermometers with iridium vessels can be applied up to 2000 °C. Thermoelements made of tungsten and a tungsten-molybdenum alloy can operate up to 2600 °C.

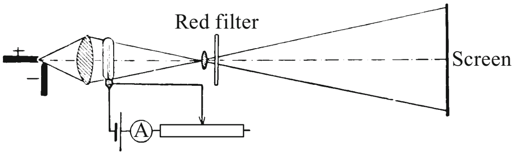


Figure 28.8 Optical temperature measurement using a pyrometer. In this demonstration experiment, the radiance of an arc-lamp condenser is compared with that of a tungsten-filament incandescent lamp. At the correct value of the lamp current, the filament of the tungsten lamp becomes invisible.

In such an application, we want to compare the radiance $L_{e,\lambda}$ within a narrow spectral range of a body of unknown temperature with the radiance $L_{e,\lambda}$ of a black body at a known temperature T. The simplest method for all comparisons is a *null method*: We vary the known temperature of the black body until its radiance is the same as that of the body whose temperature we wish to measure. Then we define the known true temperature T of the black body as the "black-body temperature" T_b of the body of unknown temperature. The black-body temperature T_b of a body thus means that within a limited spectral range *which should always be quoted*, the body radiates with the same radiance as a black body at the true temperature T. The true temperature of a body must always be higher than its black-body temperature. Otherwise, the body could not emit the same radiance $L_{e,\lambda}$ as a black body with $A_\lambda = 1$, in spite of its absorbance $A_\lambda < 1$.

Based on this definition, practical and convenient *pyrometers* are constructed. Their main component consists of a tungsten-filament lamp with variable current, an ammeter and a red filter. The tungsten filament is placed in front of the image of a radiating surface and its radiance is varied by changing the current. When the radiance of the filament and of the surface coincide, the filament becomes invisible (demonstration experiment in Fig. 28.8). The instrument is calibrated using the surface of a black body, and then the *true* temperatures of the black body are marked on the scale of the ammeter.

The discrepancies between the "black-body temperature" and the "true" temperature are often considerable, even for materials with little selective absorbance, e.g. tungsten, which is technically very important. This is shown in Table 28.1.

Table 28.1 Optical temperature measurements (in K) with tungsten

True temperature T of the tungsten	1000	1500	2000	3000
Black-body temperature T_b, measured with the radiance $L_{e,\lambda}$ in the region around $\lambda = 665\,\mathrm{nm}$	964	1420	1857	2673
Color temperature	1006	1517	2033	3094

(The ratio of the true to the black-body temperature is not constant, because the absorbance of the metal varies with temperature.)

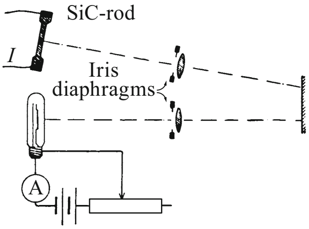


Figure 28.9 A demonstration experiment for the measurement of the color temperature. The body of unknown temperature here is an electrically-heated SiC rod. As a comparison object, a black body should in fact be used; but for this demonstration, a tungsten-ribbon lamp with a variable current source is quite sufficient. A definite color comparison requires approximately the same irradiation intensities on the screen; they are obtained by adjusting the iris diaphragms

For this reason, in addition to the black-body temperature, another temperature has been defined, called the *color temperature*. For its definition, the *unresolved* visible radiation is used, that is without a color filter, and instead of the radiant *fluxes* of the two objects, one compares their hues (red, yellow-red, etc.). Here, also, a null method is again the simplest, that is adjustment to *color equality*. A demonstration experiment is sketched in Fig. 28.9. The *true* temperature of the black body which applies at color equality is defined as the *color* temperature of the body which is being compared to it. The color temperature deviates in general much less from the true temperature than the black-body temperature does. An example is also given in Table 28.1.

Rationale: In Fig. 28.10, two solid curves of the spectral radiance $\partial L_e/\partial\lambda$ are shown for the visible range; both hold for the same arbitrarily-chosen temperature. For both, the absorbance over the whole visible spectrum has been assumed to be constant. For the upper curve, we have set $A_\lambda = 1$; it thus applies to a black body. For the lower curve, $A_\lambda = 0.6$ was chosen. The ordinates of the two curves thus differ only by a constant factor of 0.6 (bodies with an absorbance which is independent of λ, with $A_\lambda < 1$, are frequently called "grey"). The ratio

$$\frac{\text{Spectral radiance in the wavelength range around } \lambda_1}{\text{Spectral radiance in the wavelength range around } \lambda_2} = F \qquad (28.10)$$

characterizes the temperature range used (Eq. (28.5)). For our sensory perceptions of light, the *ratio F* determines the *hue* of the emitting body. The hue is thus the same for the black body and the non-black body, in spite of their different radiant *fluxes*, and conversely, the same color hue implies strictly equal true temperatures.

In general, however, the case $A_\lambda = \text{const}$ is not fulfilled for the non-black body. The lower curve then has a shape such as shown for example by the dashed or the dotted curves (Fig. 28.10). In that case, color equality means only approximate equality of the temperatures. The color temperature is lower than the true temperature in the case of the dashed curve, and it is higher in the case of the dotted curve. But such deviations become serious only for bodies with extremely selective absorption.

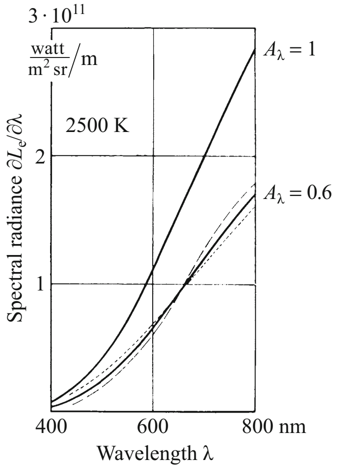


Figure 28.10 Measuring the color temperature. The values shown apply to the direction perpendicular to the emitter surface.

The blue of the sky corresponds to a color temperature of around 12 000 K; in April and May, even up to 27 000 K. This means that in the visible spectral range, the distribution of the radiance of the light from the sky, which results from diffuse (RAYLEIGH) scattering, is the same as that from some extremely hot stars (e.g. Sirius, 11 200 K, or β Centauri, 21 000 K).

Visual Perception and Photometry

29

29.1 Preliminary Remark. The Need for Photometry

The eye, like our other sensory organs, has been the object of much physiological and psychological research. Nevertheless, physicists must also be aware of the more important properties of their sense of visual perception.

In physics, we classify radiation according to its radiant *power* $\dot{W}$. (We also call this quantity the radiant *flux* to emphasize the "flow" of energy carried by a beam of radiation, e.g. towards a receiver; see Chap. 19). Figure 15.4 showed the measurement of radiant power in one of the usual units, e.g. in watt. The radiant power $\mathrm{d}\dot{W}$ as defined there is contained within a solid angle $\mathrm{d}\Omega$. Then we define the

$$\text{Radiant intensity } I_\vartheta = \frac{\text{Radiant power } \mathrm{d}\dot{W}_\vartheta}{\text{Solid angle } \mathrm{d}\Omega}\,. \qquad (19.2)$$

The radiant intensity (or simply *intensity*) is thus measured in physics as a *derived* quantity with the unit 1 W/sr.

For the sense of visual perception, the physical radiant power and the quantities derived from it (Chap. 19) are not relevant. Our visual sense responds to radiant power only very selectively in a small region of the electromagnetic spectrum. Therefore, a method of measuring the radiation had to be found in which the radiant power is evaluated only in terms of its effect on the human eye, i.e. on our visual perceptions (*photometry*). The fundamentals of photometry will be treated in Sects. 29.2 through 29.7.

Everything which is seen by our eyes, including our own bodies, consists of colored, variegated or monochrome surfaces. We see them, usually in three dimensions, as more or less bright and often shiny. Sections 29.8 to 29.13 are intended to show under which conditions our *perceptions of color and brilliance* occur.

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4_29

29.2 Experimental Aids for Varying the Irradiance

For the demonstrations in this chapter, we will require rapid and convenient variations of the *irradiance* (previously known as the irradiation intensity):

$$\text{Irradiance } E_\mathrm{e} = \frac{\text{Radiant intensity } I_\vartheta \text{ of the light source}}{(\text{Distance } R \text{ to the source})^2} = \frac{\mathrm{d}\dot{W}_\vartheta}{\mathrm{d}A'} \,. \tag{19.4}$$

This defining equation (cf. Eq. (19.4) and Comment C19.4.) shows the two possibilities: Either we change the distance R from the source to the irradiated area $\mathrm{d}A'$ (at normal incidence) in the denominator, or we change the radiant intensity I_ϑ of the source in the numerator. Among the available experimental possibilities, two will suffice for the following:

1. A rotating sector wheel ("chopper"), as shown in Fig. 29.1. It changes only the time average of the radiant intensity, and thus is completely independent of the spectral range employed. A schematic drawing of this method can be seen in Fig. 29.2 (α).

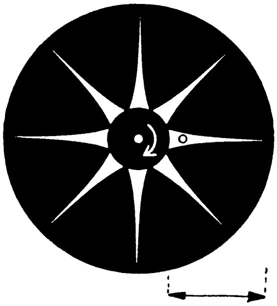


Figure 29.1 Rotating sector wheel ("chopper") for varying the time average of the radiant intensity. More than about 30 to 60 dark phases per second are no longer perceived by the eye (cf. motion picture camera!). The small black circle indicates the cross-section of the light beam. A carriage allows the chopper to be shifted to the side in the direction of the double arrows.

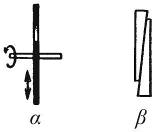


Figure 29.2 Schematic drawings of two groups of technical methods for varying radiant intensities. Left a rotating sector wheel, right two polarizing prisms.

2. Grey filters (see Fig. 27.8) of variable thickness, or two polarizing prisms or foils, one behind the other (see Sect. 24.3). They can be used only within limited spectral ranges. A schematic of this method is also given in Fig. 29.2 (β).

29.3 The Principles of Photometry

The fundamental principle of photometry is simple: We measure the radiant intensity *using our visual perception sense* as a *new base quantity*, *termed the* "luminous intensity". The unit of this new base quantity is realized in terms of the intensity of an internationally agreed-upon *standardized light source* and is called 1 candela[1] (abbreviated cd).

The meaning of this statement is explained by Fig. 29.3. Its upper part shows an incandescent lamp whose luminous intensity is to be measured; its lower part shows three standard lamps, sketched for simplicity as candles. A section of the same printed text is pasted onto the two areas dA'. The number of standard lamps has been empirically chosen so that the lower area appears to be "illuminated" by them in just the same way as the upper area is "illuminated" by the incandescent lamp: This means that one can *read the text* equally well in each of the two areas. Then *for our eyes*, we could *replace* the incandescent lamp by three standard lamps at its position and would see the text equally clearly. If each standard lamp has a luminous intensity of 1 cd, then the luminous intensity of the incandescent lamp is thus equal to 3 cd.

For the luminous intensity as a new base quantity, we then find the following juxtaposition of photometrically- and physically-determined quantities:

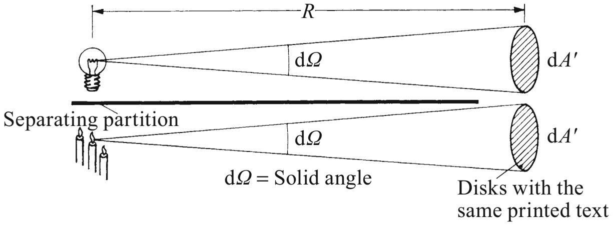


Figure 29.3 An example of the principle of photometry: *For the eye*, the incandescent lamp could be replaced by three standard lamps at the same position, represented here as candles. In the technical version, one employs only a *single* standard lamp and reduces the luminous intensity from the incandescent lamp on the upper disk to one third. The method of accomplishing this was described in Sect. 29.2.

[1] Candela (the second syllable is stressed) is the Latin word for candle. The luminous intensity is thus denoted by the same word that refers to a commercially-available object, for example a light from burning wax.

Table 29.1 Some quantities relevant to photometry, and their SI units

	Quantity	Definition	Unit
For the Source	Luminous intensity	Base quantity	candela (cd)
	Luminance	Luminous intensity/ Apparent source area (Fig. 19.3)	$\frac{\text{candela}}{\text{meter}^2}$
	Luminous flux	Luminous intensity · Solid angle	candela · steradian = 1 lumen (lm)
For the Receiver	Illuminance	$\frac{\text{Luminous flux}}{\text{Receiver area}} = \frac{\text{Luminous intensity of the source}}{(\text{Distance to the source})^2}$	$\frac{\text{candela} \cdot \text{steradian}}{\text{meter}^2} = \frac{\text{lumen}}{\text{meter}^2} = 1\,\text{lux (lx)}$

Note: The luminance is also sometimes called the "luminous density", and the illuminance is sometimes called the "illumination intensity". See Comment C19.2. and the reference there.

1. From the *source*:

Luminous intensity instead of *Radiant intensity* (Power/Solid angle)

Luminance instead of *Radiance* $\left(\dfrac{\text{Power/Solid angle}}{\text{Source area d}A}\right)$

$$\left.\begin{array}{l}\textit{Luminous flux}\text{ or}\\ \textit{Luminous power}\end{array}\right\}\ \text{instead of}\ \left\{\begin{array}{l}\textit{Radiant flux}\text{ or}\\ \textit{Radiant power}\end{array}\right\}\ \text{(Power)}$$

2. For the *receiver*:

Illuminance instead of *Irradiance* (Power/Receiver area dA')
(*Illumination intensity*) (*Irradiation intensity*)

3. For both the *source* and for the *receiver*:

Luminous energy instead of *Radiant energy*.

Up to 1979, the standard lamps were "black bodies" with an aperture of $(1/60)\,\text{cm}^2$ and a temperature of 1770 °C, the solidification temperature of platinum. Previously, before 1942, a gas lamp named for F. HEFNER (an Austrian physicist, 1845–1904) was used. Its radiant intensity in the horizontal direction was termed a "Hefner candle". One Hefner candle is ≈ 0.9 candela.[C29.1]

C29.1. Since 1979, the following definition of the base unit *candela* has been in effect: "The candela is the luminous intensity in a particular direction of a radiation source which emits monochromatic radiation at the frequency 540 THz and whose radiant intensity in that direction is (1/683) W/sr"; see e.g.: *PTB-Mitteilungen* **117** (2007), No. 2; English: see http://physics.nist.gov/cuu/Units/current.html Specially-constructed incandescent lamps are employed in practice as secondary standards.

The *base quantity* 'luminous intensity' *suffices in order to determine all the other quantities needed in photometry as derived quantities*[2]. If the units of these derived quantities are given special *names*, as is often the case, then this harmless measurement specification takes on the aspect of a truly esoteric doctrine. In Table 29.1, we have collected the names of some of these quantities and their units.

[2] Of course, one could introduce other physical quantities which can be related to our visual perception sense as base quantities, for example the luminous density or photometric brightness (now called the "*luminance*"), i.e. the radiance as *perceived by the eye*. Then the luminous intensity would become the derived quantity luminance · source area, etc. The use of the luminous intensity as base quantity makes it *experimentally* simpler to develop the photometric measurement procedures.

29.4 Definition of the Equality of Two Illuminances

All of photometry stands and falls with our ability to identify two areas or "fields" which are *perceived* as "equally illuminated" by different light sources; or, more precisely stated, we perceive their illumination intensities, or *illuminances*, as being equal. In comparing two light sources of similar construction, for example a large and a small tungsten-filament incandescent lamp under normal current strengths, the approach to equal illuminances is readily detected. We let the two areas dA' in the schematic of Fig. 29.3 be adjacent to each other. At a precisely equal illuminance, the *boundary* between them *vanishes*, and we can no longer distinguish the two illuminated areas from each other.

The situation is different in the *comparison of different types of light sources*, e.g. a yellow sodium-vapor lamp and a blue-green mercury-vapor lamp, or two arc lamps with colored filters F_1 and F_2, one a red filter and the other a blue filter. In such cases, the concept of equal illuminance or illumination intensity must first be *defined*. There are several possibilities for accomplishing this:

1. *Visual sharpness* or acuity. This possibility was already explored in Sect. 29.3. Now, we suppose that on a sheet of newsprint, there are two rectangular fields next to each other. Each one is illuminated with an arc light, one with a red filter, the other with a green filter (Fig. 29.4). The illuminance of the one field can be continuously varied in a quantitatively measurable manner using the apparatus β. With a remarkable accuracy, one can adjust the two fields to have the same *legibility* or the same *visual sharpness*. Therefore, for any color, we can *define* the visual sharpness or acuity as characterizing equal illuminances.

2. *Delay time*. The two rectangular, colored fields are projected onto a wall screen side by side with a vertical boundary between them, but interrupted by the shadow of a horizontal rod. The rod is moved up and down. Its moving shadow in general doesn't appear to be a horizontal straight line, but instead it seems to show a dislocation at the boundary between the two fields, as shown e.g. in Fig. 29.5. That is, our sensory system perceives the motion only with a certain *delay*

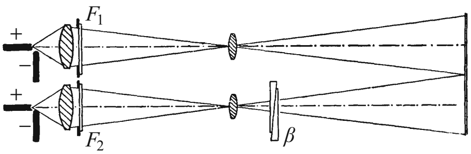
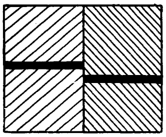


Figure 29.4 The definition of equal illuminances using visual acuity (sharpness). The surrounding area in this and in the following photometric demonstration experiments is lit with an illuminance of about 10 cd/m^2. Then it radiates diffuse light itself, with a luminance of about 3 cd/m^2.

Figure 29.5 The definition of equal illuminances through equal delay times



time, which is dependent on the illumination intensity. Again, we can vary the illuminance of one of the fields continuously (apparatus β), and then with good accuracy, we can adjust it to eliminate the apparent dislocation of the rod's shadow. Thus, independently of the color, we can use the equal delay times to *define* equal illuminances.

In technical photometers, stereoscopic effects can be produced by delays of differing lengths. When they disappear, the illuminance is equal. *Demonstration experiment*: Hang a metal ball by two cords (bifilar, length ≈ 4 m) as a gravity pendulum and let it swing *in one plane* (the bifilar suspension guarantees this; $T \approx 4$ s). The observer views it from the side with both eyes, but holds a piece of dark or colored glass in front of one eye. Then the ball appears to follow an elliptical path. The direction in which it moves around the ellipse depends on whether the reaction time of the left eye or the right eye is delayed by the dark glass.

3. The *limiting frequency of flickering*. Intermittent illumination (stroboscopic light), produced e.g. by a rotating sector disk with radial sectors as in Fig. 29.6, causes flickering. This flickering appears to vanish above a certain *limiting frequency*[3]. The higher the illumination intensity (variable using the apparatus β), the higher the limiting frequency. When the illumination is of different colors, the same limiting frequency of flickering can be *defined* to indicate equal illuminance.

4. *Flicker-free field exchange*. The two fields illuminated with colored light are no longer projected *beside* each other, but are rather precisely *superimposed* on each other (Fig. 29.7) and presented *alternately* to the eye of the observer, around 10 times per second. In

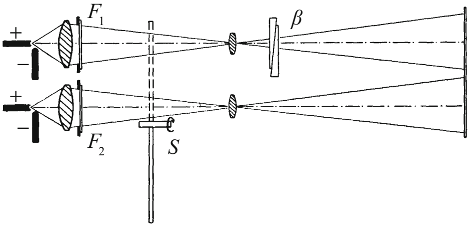


Figure 29.6 The definition of equal illuminances by equal *limiting frequencies* of flickering. The rotating sector disk S with radial sectors opens both light sources simultaneously and for equally long times.

[3] The limiting frequency is, from experience, smallest when the durations of the dark and light intervals are *equal*. (In motion-picture films, about 0.01 s. Every image is projected twice and only every second dark interval is used to change to the next image; the image frequency is thus 25 Hz.)

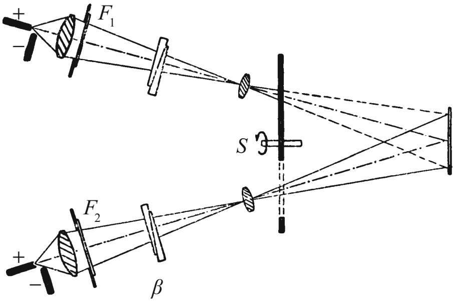


Figure 29.7 The definition of equal illuminances by means of a flicker-free field exchange

general, one sees the exchange of the two images as a flickering of the color hue. By changing the illuminance of one of the sources (apparatus β), one can suppress the flickering. The eye then sees the fields in an unchanging, mixed color. This flicker-free field exchange can be *defined* as an indicator of equal illuminances, independently of the colors used.

These different definitions for the equality of two illuminances lead to moderately good agreement among the results[4]. Making use of them, the luminous intensities of various sorts of light sources can be compared and measured in multiples of the conventional unit, the candela. *The numerical values obtained from photometry naturally hold only for an "average normal human"; and even for such a person, they hold only under normal conditions, and not when some sort of stress-related disturbances have affected the subjective well-being of the person.*

29.5 The Spectral Distribution of the Sensitivity of the Eye: The Luminosity Function

According to the considerations in the previous section, luminous intensities *for any color* can be measured in candela. As a result, we can determine experimentally how the ratio[C29.2]

$$E_\lambda = \frac{\textit{Luminous}\text{ intensity, measured photometrically in candela}}{\textit{Radiant}\text{ intensity, measured physically in watt/steradian}} \tag{29.1}$$

C29.2. It is equivalent to the definition

$$E_\lambda = \frac{\text{Luminous flux}}{\text{Radiant flux}}$$

(unit: lumen/watt), as plotted on the right ordinate axis of Fig. 29.8.

[4] One would have to give preference to the definition whose results best obey an additivity rule. Illuminances are additive; for example, following one of the methods described, we determine two illuminances A and B. When added, they yield the illuminance $C = A + B$. When a direct measurement by the same method also gives the value C, we can say that the method obeys the additivity rule. In this sense, definition no. 4 appears to be the best.

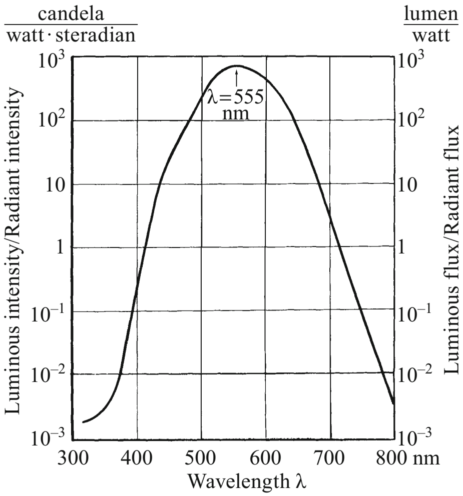


Figure 29.8 The spectral distribution of the luminosity function or luminous efficiency E_λ of the eye, for an eye adapted to bright light, as defined by the current internationally agreed-upon values.[C29.3] If the 10 % of all male observers with minor color-vision disturbances are eliminated, then the maximum would be shifted to a wavelength of 565 nm. Conventionally, the wavelength range from 400 to 750 nm is termed "visible". This is, however, somewhat arbitrary.

C29.3. The measurement of the luminous efficiency curve (Fig. 29.8) is a complex process. Here, as in Sect. 28.3, we are dealing with wavelength *intervals* which are small and are denoted by their central wavelength. The curve was established by the International Commission on Illumination (*Commission Internationale de l'Eclairage*, CIE) in 1931 for a bright-adapted eye ("photopic curve"). See e.g. https://en.wikipedia.org/wiki/Luminosity_function and http://www.cie.co.at/ for more details.

depends on the wavelength of the radiation. One can denote E_λ as the *spectral sensitivity* of the eye, or alternatively as the *luminosity function* or *luminous efficiency* as a function of the wavelength λ.

The methods of measurement are sufficiently well known to us from the previous sections. The result – averaged over a year for hundreds of individuals – is plotted in Fig. 29.8. It holds for an eye which is adapted to bright light, i.e. for a state of the eye which applies when the luminance of light sources (lamps) or the illuminance of reflected light (room walls, furniture, printed pages) is $> 3\,\mathrm{cd/m^2}$. The maximum of the curve then lies at a wavelength of $\lambda = 555$ nm; at this wavelength, $E_{\max} = 683.002\,\frac{\mathrm{cd}}{\mathrm{W/sr}} = 683.002\,\frac{\mathrm{lm}}{\mathrm{W}}$.

The position of the maximum of the spectral sensitivity of the eye can be demonstrated qualitatively by very simple experiments. We project the spectrum from an arc light onto a wall screen and assume that the radiant intensity of the different wavelength regions is roughly constant, a sufficiently good approximation. We then place a sector disk in the optical path and gradually increase its rotational frequency. At first, the whole spectrum flickers, then its ends (violet and red) become flicker-free. The region which is still flickering becomes narrower and narrower. *Finally, the limiting frequency of flickering is reached in the green region of the spectrum; this is the region of highest sensitivity*. Or, still more simply: We remove the sector disk and hold a needle horizontally in front of the slit. It divides the spectrum along its whole length horizontally by a straight, dark streak. We then move the needle slowly up and down; the dark streak appears to be bent into a curved arc, its ends in the red and violet regions lag behind. The top of the arc is in the green region, i.e. again in the region of maximum sensitivity, where the delay time of the eye is shortest.

At low illumination intensities, the light receptors of the bright-adapted retina of the eye, the "cones", are no longer functional. Instead, other receptors, the "rods", take over. At illumination intensities of $< 3 \cdot 10^{-3}\,\text{cd/m}^2$, only the rods are active. The spectral sensitivity distribution of the eye is then shifted towards shorter wavelengths. Its maximum lies around 510 nm. The eye still reacts to an irradiance of around $6 \cdot 10^{-13}\,\text{W/m}^2$, i.e. its pupil, with an area of $5 \cdot 10^{-5}\,\text{m}^2$, must pass a radiant power of about $3 \cdot 10^{-17}\,\text{W}$ or a luminous flux of around $2 \cdot 10^{-14}\,\text{lm}$[5]. With the rods alone, the eye can no longer see objects in color – "At night, all the cats are grey". The rods are not present in the angular region of greatest visual acuity (see the last paragraph of Sect. 18.14). Therefore, objects seem to disappear when looked at directly (fixed), and they reappear when one looks past them. We see "ghostly lights" and will-o'-the-wisps.

"At night, all the cats are grey".

To demonstrate these effects, we completely darken the lecture hall and project a spectrum onto the wall screen while varying the illumination intensity from the slit using two crossed NICOL prisms or Polaroid filters (Sect. 24.3). After a few minutes, the observers are all dark-adapted. The spectrum appears to be a silvery-shining band, with a clear maximum intensity in the region that was previously "blue". When one fixes an object directly, it is invisible; one has to "look past it" in order to see it.

Having determined the two spectral sensitivity distributions for the bright- and the dark-adapted eye, we have laid down the physiological basis of photometry. For technical and economic purposes, aptly-chosen average values (e.g. as in Fig. 29.8) can be agreed upon as internationally binding. Based upon them, all the necessary practical measurements of light can be carried out with *instruments* only, without reference to human visual perceptions. It is not difficult to provide a photoelectric radiometer (a photocell and an ammeter, see Fig. 15.6) with the same spectral sensitivity as that of the eye. The selective photoelectric effect of the alkali metals, especially cesium, is very well suited for this purpose (cf. 13th edition of "*Optik und Atomphysik*", Chap. 18); it can be combined with special filters. Such arrangements are often called *objective photometers*. They indicate the radiant power (watt) with the same measure, *which varies with wavelength*, as a conventional, average "normal eye". The scale of the ammeter can be directly calibrated in a photometric unit, e.g. candela. In this and other forms, technical photometry solves by convention the task of providing economically usable data and avoiding controversies. For the vision of a single individual, these data are of course not strictly applicable. Where the results of such data contradict visual perceptions, the eye is always right!

"Where the results of such data contradict visual perceptions, the eye is always right!"

[5] Corresponding to roughly 100 light quanta/second.

29.6 The Rise Time and Accumulation Time of the Eye

The results of the previous section hold only for steady illumination of the eye. Only then is the luminous intensity proportional to the *radiant power*. Our visual sensory perceptions are initiated by photochemical processes in the retina of the eye. The concentration of their reaction products is by no means always and continuously proportional to the absorbed light *energy*. Thermal processes and biological regeneration in the living cells effect a reconstitution. As a result, the concentration never exceeds a steady-state limiting value which is proportional to the radiant power. This value is however reached only after a certain *rise time*. As long as the irradiation time is short compared to this rise time, the photochemical reaction products are accumulated. During the accumulation time, only the *product* of the radiant power and the irradiation time is relevant; it measures the *energy* input.

Example: For the bright-adapted eye, the accumulation time is $\tau \approx 0.05\,\mathrm{s}$. Therefore, one can look at the disk of the sun (luminance $\approx 10^9\,\mathrm{cd/m^2}$) for a time of e.g. $5 \cdot 10^{-5}\,\mathrm{s}$ without perceiving it to be brighter than a continuously-observed, weakly glowing tungsten ribbon lamp (luminance $\approx 10^6\,\mathrm{cd/m^2}$).

Application: We can allow a lamp to glow continuously, and then by means of brief overload pulses, we can produce additional flashes of light, for example to serve as signals. The human eye does not perceive the signals, but they can be registered by a receiver which makes use of a photo detector with a very short rise time.

29.7 Brightness

This often-used word from everyday language is rather ambiguous. It denotes for example the quality of a sensation: the chromatic hue ‘violet’ can never be perceived to be as bright as the chromatic hue ‘yellow’. Usually, brightness is used in the sense of *luminance*, measured in candela/m^2, both for primary light sources and for reflection (secondary) sources. In addition, everyday language uses the word ‘brightness’ for the *luminous intensity* of a lamp, of a firefly, etc., measured in candela, without considering the *size* of its radiant area. In *astronomy*, finally, the word ‘brightness’ is used in three different ways; most frequently in the sense of *illuminance*:

$$E_\mathrm{L} = \frac{\text{Luminous flux}}{\text{Receiver area}} = \frac{\text{Luminous intensity } I_\mathrm{L} \text{ of the star}}{(\text{Distance } R \text{ to the star})^2}\,. \quad (29.2)$$

Astronomers compare only the illuminances $E_{\mathrm{L},1}$ and $E_{\mathrm{L},2}$ from two stars as registered on the earth. Then they define (on the basis of a long historical development) the *visual magnitudes* m_i by the equation

$$m_2 - m_1 = 2.5 \log \frac{E_{L,1}}{E_{L,2}} . \tag{29.3}$$

The difference between two magnitudes m_2 and m_1 is thus proportional to the logarithm of the ratio of the corresponding illuminances from the two stars. The value m_1 is taken to be $+2.12$ for the north polar star, Polaris (an arbitrary convention). On this scale, the visual magnitude m_2 of a star which can just be discerned by the naked eye is equal to $+6$. Its value for α Cygni (Deneb) is $+1.3$; for Sirius, it is -1.6; and for the sun, it is -26.7. (Compare the definition of the *phon* in Vol. 1, Sect. 12.29).
In Eq. (29.3), the illuminance E_L is used. With known distances R, astronomers instead use the *luminous intensity* $I_L = E_L R^2$ (Eq. (29.2)), and define the difference between two numbers M_1 and M_2 by the equation

$$M_2 - M_1 = 2.5 \log \frac{I_{L,1}}{I_{L,2}} = 2.5 \log \frac{E_{L,1} R_1^2}{E_{L,2} R_2^2} . \tag{29.4}$$

These numbers are called the *absolute* magnitudes or *absolute brightness*. Combining Eqns. (29.3) and (29.4) yields

$$M_2 - M_1 = m_2 - m_1 + 5 \log \frac{R_1}{R_2} . \tag{29.5}$$

For a fixed star at a distance of $R_1 = 10$ parsec[6] which has a "visual magnitude" of $m_1 = 0$, M_1 is assigned the value 0. Thus, for a fixed star at a distance R_2 and with a visual magnitude of m_2, we find the *absolute magnitude* to be given by the *number*

$$M_2 = m_2 + 5 \log \frac{10\,\text{parsec}}{R_2} , \tag{29.7}$$

or, if we use Eq. (29.9) to express its distance R_2 as a parallax[7] from the earth, replacing the distance by α_2:

$$M_2 = m_2 + 5 + 5 \log \frac{\alpha_2}{1''} . \tag{29.10}$$

The brightness or magnitudes which are termed 'absolute' are thus those which would be observed from a distance of 10 parsec.

[6] In astronomy, the distance unit 'parsec' is equal to that distance R_0 from which the radius of the earth's orbit r would be seen to subtend an angle of $1''$, that is

$$R_0 = r/1'' = 1\,\text{parsec} = 3.08 \cdot 10^{16}\,\text{meter} \tag{29.6}$$

$$(1'' = (1/3600)^\circ = 4.85 \cdot 10^{-6}\,\text{rad}, r = 1.49 \cdot 10^{11}\,\text{m}).$$

[7] The *parallax* α of a fixed star is defined as the angle

$$\alpha = \frac{\text{Radius of earth's orbit } r}{\text{Distance to the fixed star } R} . \tag{29.8}$$

From Eqns. (29.6) and (29.8), we find that a fixed star with a parallax α is at a distance

$$R = \frac{1''}{\alpha} \cdot R_0 = \frac{1''}{\alpha}\,\text{parsec} \tag{29.9}$$

Table 29.2 Examples of luminances

	Luminance in cd/m^2
Primary light sources	
Night sky	ca. 10^{-11}
Neon lamp	ca. 10^{-5}
Gas mantle lamp	$6 \cdot 10^{-4}$
Mercury arc light	$(2–6) \cdot 10^6$
Tungsten filament lamp (gas-filled)	$(5–35) \cdot 10^6$
Carbon arc crater (black-body temperature $= 3820$ K)	$1.8 \cdot 10^8$
ditto, with added cerium fluoride (BECK lamp)	$(4–12) \cdot 10^8$
High-pressure mercury arc lamp (quartz bulb, 45 bar)	up to[a] $6 \cdot 10^8$
Sun	$(10–15) \cdot 10^8$
Reflective sources (secondary sources)	
Objects in illuminated working and living rooms	$< 10^2$
Objects in workplaces for very fine or precision work	ca. 10^3
Objects on the street, sun behind the observer	ca. $5 \cdot 10^3$
Objects outside in cloudy weather	ca. $3 \cdot 10^3$

[a] For short times (fractions of a second), this maximum value may be greatly exceeded, up to a large multiple of the luminance of the sun.

The *photographic brightness* refers to the ratio

$$E'_\mathrm{e} = \frac{\text{Photochemically-evaluated radiant intensity } I \text{ of the star}}{(\text{Distance } R \text{ of the star from the earth})^2} . \quad (29.11)$$

Using the quantities E'_e instead of E_L, the astronomers then define *numbers*, using an equation analogous to Eq. (29.3), which are termed "photographic magnitudes".

The *luminosity* in astronomy refers to two different quantities: First, the total *radiant power* emitted by a star (measured e.g. in watt); and second, a *pure number*, namely the ratio of the absolute magnitude of a star to the absolute magnitude of the sun ($M_\odot = +4.8$). These numbers lie between around 10^5 (for giant stars) and 10^{-6} (for dwarf stars).

"Given this desolate muddle, we should avoid the word 'brightness' as far as possible".

Given this desolate muddle, we should avoid the word 'brightness' as far as possible. The

$$\text{Luminance } L_\mathrm{L} = \frac{\text{Luminous intensity}}{\text{Projected source area (Fig. 19.3)}} \quad (29.12)$$

is independent of the direction of emission when LAMBERT's cosine law applies (Sect. 19.2); that is, both for primary light sources and for ideally diffuse-scattering *reflectors*. Therefore, we see the sun's *sphere* as if it were a uniformly-emitting *disk*, just like a white ball that is illuminated uniformly from all sides (see however the footnote in Sect. 19.3).

The eye can *adapt* itself to an astonishingly wide range of luminances, namely the range between $2 \cdot 10^{-6}$ and $2 \cdot 10^5$ cd/m^2. For every state of adaptation, a certain maximum luminance should not be exceeded; otherwise *dazzling* or temporary blindness occurs, i.e. the visual acuity of the eye and its ability to distinguish colors are greatly reduced. At the upper limit of the eye's adaptation range, first discomfort and then pain warn of the impending permanent eye damage. The luminances of many light sources exceed the adaptation range of the human eye. This is shown in Table 29.2.

The influence of optical instruments such as telescopes on the luminance of the objects being observed is very important. But here, psychological aspects also play a significant role.

29.8 Achromatic Colors. The Conditions for Their Formation

The achromatic colors can be arranged along a series, called the *grey scale*. At one end is white, at the other end is black. The transition between them includes all the grey colors. It can be continuous or in discrete steps, e.g. in 10 steps of equal change in contrast. To demonstrate the grey scale, we first make use of commercially-available white, grey and black paper (Fig. A1 on Plate 1 at the end of this chapter) and illuminate it with *sunlight* or with *natural light* from an incandescent lamp (see the end of Sect. 16.1).

Physically, all these achromatic surfaces have one thing in common: In the visible spectral range, their reflection coefficients depend only weakly – in the ideal case not at all – on the wavelength of the light. The different achromatic surfaces differ only in their overall reflection coefficients. This is around 90 % for white paper, but only about 6 % for black paper. For this reason, all achromatic surfaces (white, grey, and black) reflect visible radiation with the same spectral distribution; only their *radiance* differs: $(\frac{\text{Power}}{\text{Solid angle}\cdot\text{Area}}$, unit: $\frac{\text{W}}{\text{sr}\cdot\text{m}^2})$, or, measured photometrically, their luminance (in cd/m^2). This corresponds to a very important empirical fact: *Every achromatic surface appears to have the same color*, white, *when it is illuminated by itself in a darkened room with natural light*.

Demonstration experiment: In Fig. 29.9, a circular aperture is placed in front of the condenser lens of an arc lamp and imaged on a screen. In the well-lighted lecture room, a white cardboard sheet is hung up, and then the lights are dimmed: One then sees a luminous white circular disk. Then the arc lamp is switched off, the white cardboard is secretly replaced by a black cardboard, the power to the arc lamp is increased and the observation is repeated. The observer again sees a luminous white disk, something like the *luminous white moon against the dark sky*.

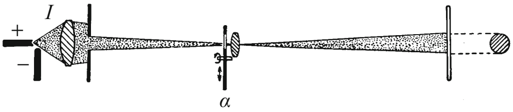


Figure 29.9 How the color white comes about (**Video 29.1**)

Video 29.1:
"The color white"
http://tiny.cc/dgggoy

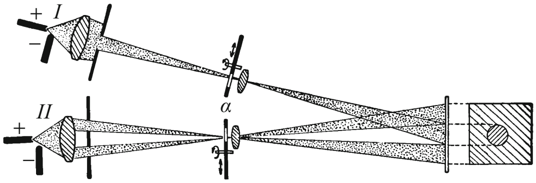


Figure 29.10 How the colors grey and black are formed

An achromatic surface *alone* can thus never be seen as grey or black. *To see grey or black, the eye requires a second area with a higher luminance in its field of view*. This can be arranged in two different ways: Either we use at least two different achromatic surfaces with different reflection coefficients and illuminate them with a common source of natural light; or we use only *one* achromatic surface and illuminate two separate fields on it with two lamps of different luminous intensity. The latter method is shown in Fig. 29.10. Lamp *I* illuminates a circular "inner field", while lamp *II* illuminates only an "outer field" which is bounded outside by a rectangular border. Both illumination intensities can be varied continuously over a wide range using the apparatus α (e.g. sector wheels; cf. Fig. 29.1).

At first, only the inner field is illuminated and the luminance of its scattered radiation is adjusted to a medium value. The inner field appears to be pure white. Then, the inner field – without changing anything in its own illumination – is surrounded by a luminous outer field. Immediately, the color of the inner field changes to grey. The more intense the illumination of the outer field, the darker the grey of the inner field. We can pass through the entire grey scale up to deep black, without, we repeat, making any changes at all in the illumination of the inner field. At the end, the illumination of the inner field is removed, so that only its illuminated *frame* remains. Now, the inner field appears to be blacker than the best matte black paper or even soot.

Summary: A surface becomes black not through its *own* radiation, but rather as a result of the radiation from its surroundings. Without light, we see nothing at all; black can be seen only through light from the surroundings. Grey colors are formed by the presence of *two* visible radiations. One of them is radiated from the *object* itself, the other from its *surroundings*. The ratio of the two luminances determines the degree of greyness (the grey scale value).

29.9 Chromatic Colors, Their Hues, and Their Tinting and Shading

With natural light, we produce a continuous spectrum. Even when we look at it only fleetingly, we are surprised by the *lack of variety* of different colors. A large group, the purple hues (wine red, etc.), is

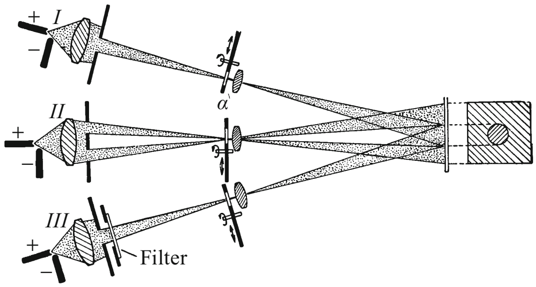


Figure 29.11 Tints and shades of chromatic colors (**Video 29.2**)

Video 29.2:
"Tinting and shading of chromatic colors"
http://tiny.cc/sgggoy

completely missing. We search in vain for the most frequently-used colors of our clothing, our furniture and our wall coverings. There is no brown, no pink, no dark green, etc. *Alongside a commercial color chart, the range of colors in the spectrum appears limited and paltry*.

We can, to be sure, quite properly refer to radiation within a very narrow wavelength range as *monochromatic*, that is as a 'pure color': *We see it as a characteristic chromatic hue* or 'chroma'. But we cannot turn this sentence into its converse: *Only in rare cases is monochromatic light*, that is light that we see as a 'pure' chromatic hue, identical to the radiation from a *narrow* wavelength range! (Sect. 29.10).

It is difficult to bring order into the sheer unmanageable variety of chromatic hues. Every chromatic color exhibits a certain hue which cannot be precisely defined, namely red, yellow, purple etc. In addition to this hue, another characteristic *may* be present: The color may be *tinted* or *shaded*, i.e. a red can be whitened (tinted) or darkened (shaded); it may furthermore often show a grey shading.

'Pure' colors which are not tinted or shaded can be produced using filters and can be ordered on a circle according to their similarities (cf. Fig. A2 in Plate 1 at the end of this chapter). In this *color circle*, every hue is more similar to its two immediate neighbors than to any other hue on the circle.

In order to *demonstrate tints and shades*, Fig. 29.10 is complemented by a third projection lamp (Fig. 29.11). It illuminates only the *inner field*, initially through a red filter. Then four experiments are carried out, one after the other:

1. Only lamp *III* is turned on; the inner field appears as a pure or *free* red hue.

2. Now we want to add a *white tint* to the red in the inner field. To accomplish this, the inner field is additionally illuminated with light from lamp *I*, and its illuminance is gradually increased using the apparatus α. This leads us from the pure red through pink to an achromatic white.

3. In the third experiment, the red of the inner field is to be given a *dark shading*. Lamp *I* is switched off, and in its place, the outer

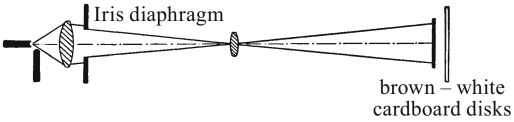


Figure 29.12 Black shading of a chromatic color is possible only with the aid of a second illuminated area

field is illuminated with increasing intensity by the natural light from lamp *II*. We pass from the pure red via pleasant dark red hues to an achromatic black in the inner field.

4. Finally, the pure red of the inner field is to be given a *grey shading*. To accomplish this, it has to be overlaid simultaneously with white and black, that is both the inner field (lamp *I*) and also the outer field (lamp *II*) illuminate the screen with their natural light. Making use of the apparatus α, we can vary the illuminance from both lamps and can pass from a pure or free red via greyish-red to any arbitrary shade of achromatic grey in the inner field.

All the tints and shades of a given color hue can be represented two-dimensionally by "HERING's triangle" (Fig. A3 on Plate 1 at the end of this chapter). One such *tinting and shading triangle* thus corresponds to each individual hue on the color circle. In this manner, all the manifold colors in nature and art can, through the right combinations, be catalogued and designated by letters and numbers. This is used in commercial color charts, which are based without exception on the classic works of EWALD HERING (1834–1918).

A black- or grey-shaded chromatic color can, just like a black or grey shade, never appear alone in the field of view. In Fig. 29.12, a brown paper disk is placed in front of an achromatic screen and, in the darkened lecture hall, it is first illuminated with natural light. One sees no brown, but instead only the corresponding unshaded hue that corresponds to brown, a reddish-yellow color.[8] Then the light beam which illuminates the disk is broadened so that the screen behind is also illuminated. Immediately, a dark shading is produced, and the reddish-yellow becomes a typical brown.

29.10 Color Filters for Producing Pure Hues

In the previous section, we saw that we could produce luminous colors free of any white tint by using suitably-chosen filters. What range $\Delta\lambda$ from the spectrum of natural light must such a filter trans-

C29.4. This disrespectful experiment was described in POHL's "*Optik*" beginning with its 1st edition in 1940. See also the addendum to this Comment at the end of this chapter.

[8] The formation of brown shades can be demonstrated by quite simple means. A circular disk has three glued-on sectors of colored paper, around 210° of black, 90° of red, and 60° of yellow. It is rotated rapidly. The motion causes the three individual colors to vanish and 'merge' together into a unified brown.[C29.4]

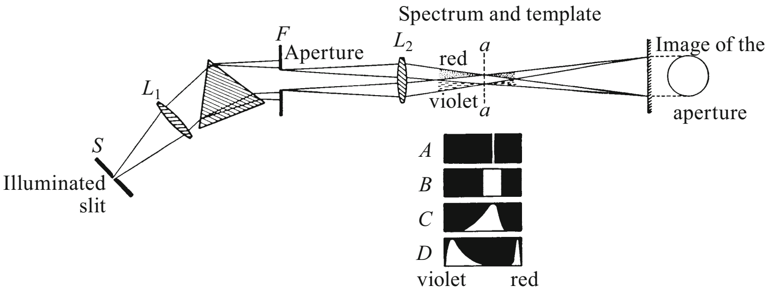


Figure 29.13 The spectral distribution of pure hues with a high luminous intensity. With a small modification, this arrangement is also suitable for the demonstration of *complementary colors*. For this, we replace the template by a narrow prism with a small angle of refraction. Then we can obtain two images of the aperture on the screen, preferably partially overlapping. These two images are colored with complementary hues; in the overlap region, an achromatic circular segment ('lune') is seen. Tilting the auxiliary prism along the spectrum allows us to image a number of different pairs of colors (**Video 29.3**).

mit? We wish to answer this question experimentally as shown in Fig. 29.13, in particular, as an example, for a pure, untinted and unshaded green hue (chroma).

The incandescent light of the arc lamp emerges from the slit S and illuminates the aperture F. On the way, it is spectrally decomposed by the prism and the lens L_1. The image of the aperture is projected onto the wall screen by the lens L_2. The continuous spectrum appears in the plane aa. In this plane, a template made of opaque cardboard is placed in the optical path. We first use a very narrow slit (A in Fig. 29.13) and locate the desired chromatic hue in the spectrum. We then widen the slit gradually on both sides until it is a broad rectangle. By trial and error, we find the largest width that is still permissible (template B); it gives the maximum possible luminous intensity without white tinting. The ratio $\Delta\lambda/\lambda$ approaches a value of several tenths, so that the light is by no means monochromatic (in the sense of 'single-frequency')! Finally, the steep sides of the template can be tapered off somewhat (C). This changes neither the luminous intensity nor the purity of the colors. In the same way as for the green hue here, we can fabricate a broad template for every other pure hue in the color circle. The template D in Fig. 29.13 shows an example for a purple hue[9].

Templates of this type, with tapered and bent flanks, can be replaced by filters containing selectively-absorbing substances. The filter must allow the same broad regions of the spectrum to pass through as does the template. For demonstration experiments, we proceed as shown in Fig. 29.14. We image the spectrum of the incandescent light (arc lamp) on the wall screen, placing a filter cuvette in front of the slit,

Video 29.3:
"Complementary colors"
http://tiny.cc/4gggoy
In the video, some parts of the spectrum are not blocked off by the template, but rather are deflected perpendicular to the spectrum by a thin quartz prism.

[9] For the purple hue, we need two narrow slits, one in the blue or in the violet region, and the other one in the red.

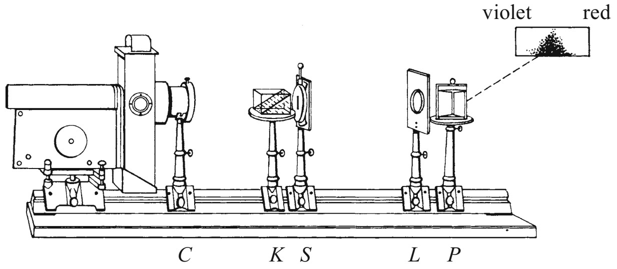


Figure 29.14 A demonstration experiment on the selective light absorption by dye solutions. The solutions are placed in the form of a wedge-shaped layer in front of the entrance slit of the spectral apparatus. A similar wedge-shaped layer of water prevents a disturbing deflection of the light due to refraction. At the upper right, we see the transmitted region of the spectrum. The densely dotted area indicates a high luminous intensity. Imagine that the spectrum has been decomposed into horizontal stripes. The uppermost stripe corresponds to the thickest layer of solution (the image is upside down!); only a narrow region of the spectrum is transmitted. A middle horizontal stripe in the spectrum corresponds to a medium layer thickness; here, the transmitted region represents several tenths of the whole spectrum; and so forth.

with a diagonal separation partition. The cuvette which is shown as transparent contains pure water; the shaded cuvette contains a dye solution. Then the spectrum appears on the screen with only one wavelength region, as if it were bounded by a template. The composition of the dye solution as well as its thickness are varied until the outlines of the spectrum are identical with the opening of the desired template.

Finally, we remove the prism, open up the slit fairly wide, and image it by itself on the screen. Behind the greatest filter thickness, it is imaged in pure chromatic hues. With decreasing filter thickness, the hues are tinted more and more with white.

Starting with a large selection of light filters, those that are suitable for producing pure, free hues can be selected with relative ease. We simply observe a color circle through each filter; those that are suitable show one *half* of the circle as bright, and the other half dark. The brightest sector of the color circle shows the hue that can be selected by that filter.

The concentration and layer thickness of color filters select by no means only the degree of white tinting, but rather often the *chromatic hue* itself. To demonstrate this, we image a tungsten-filament lamp on the screen and allow its light to pass through a wedge-shaped filter, for example containing an indigo solution, that can be slid in the longitudinal direction (wedge angle $\approx 20°$, $c \approx 0.05$ g/liter). After passing through a short filter layer thickness, the light from the lamp appears blue-green; a longer thickness gives wine red. The *color transition* occurs rather sharply at a layer thickness of about 3.5 cm for a lamp operating at the usual temperature.

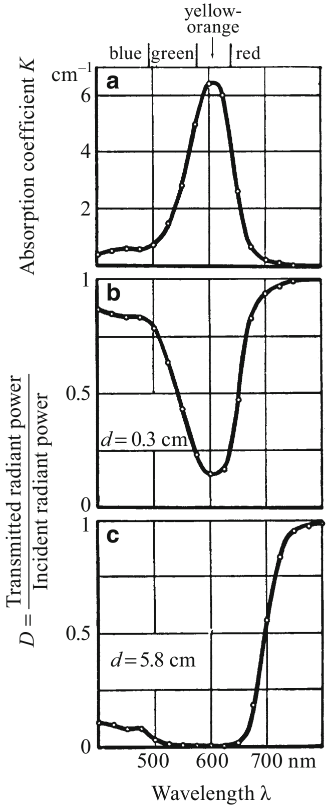


Figure 29.15 The dependence of the chromatic hue of a filter on its layer thickness d

The explanation can be seen in Fig. 29.15. Part a of the figure shows the dependence of the light absorption constant K on the wavelength. It is small in the red wavelength region, then passes through a maximum in the orange ($\lambda = 605\,\mathrm{nm}$), but still remains larger in the blue than in the red region.

Parts b and c of the figure show the transmission coefficients D, defined by the equation

$$D = \frac{\text{Transmitted radiant power}}{\text{Incident radiant power}}\,, \tag{29.13}$$

for a small and a large layer thickness. At the small layer thickness (b), the short and the longer waves are transmitted equally well: The broad region of shorter wavelength predominates and therefore the light from the lamp appears blue-green. At a larger layer thickness (c), the shorter waves are much more strongly absorbed than the longer waves; thus, the red spectral region predominates and the light from the lamp appears to be wine red.

The layer thickness which corresponds to the color transition varies with the temperature of the lamp (demonstration!). We can thus recalibrate the thickness scale as a temperature scale. This provides us with a very convenient instrument for measuring color temperatures (Sect. 28.7).

29.11 Dyes and Pigments

The filters contain selectively-absorbing substances in the form of suspended solids or in solution. Among such solutions are most of the technical dyestuff and pigment materials. These layers can be applied to an object in two ways:

In *enamel or varnish paints*, the solution is free of any sort of inhomogeneities which would make it cloudy. One can make out details of the surface through the paint layer. Light coming from a light source passes through the paint layer to the surface of the object and is scattered there. It thus passes through the whole layer thickness *twice* before reaching the eye of the observer. As a result, with the proper concentration, rather pure chromatic hues can be obtained, free of tinting or shading. The reflected light from the front surface of the painted layer may cause a weak white tinting, but this surface can be made smooth as a mirror, thus limiting this disturbance to the angular range of specular reflections. Still better is the *elimination* of surface reflections. This is most successful with textiles of the velvet type. *Black shading* (e.g. in numerous articles of clothing) can be obtained in any desired degree; one need only increase the *concentration* of the absorbing substances.

The *opaque or topcoat paints* are made artificially *cloudy*. Usually, the selectively-absorbing substance is not dissolved, but rather is suspended as a fine powder in a binding agent. The incident light cannot reach the surface of the painted object; it is scattered backwards by the fine particles. Thus, the surface of the object is covered, and an essential portion of the light passes through only some *fraction* of the paint layer thickness. As a result, there is a considerable *white tinting*. A black shading can also be obtained by using high concentrations of the absorbing substance, but the ever-present white tinting has a disturbing effect. The tinting and the shading together lead to a *grey shading*, and often, grey shadings appear rather "dirty".

"The contrast between varnish and topcoat paints can be demonstrated with any cup of tea".

The contrast between varnish and opaque paints can be demonstrated with any cup of tea. Clear tea forms a "varnish layer", and the tea leaves at the bottom of the cup are clearly visible. Adding some drops of milk converts the varnish layer into a cloudy topcoat layer. The bottom of the cup is no longer visible, and at the same time, a strong white tinting can be observed.

In general, the colors of objects, both natural colors and those produced by pigment coatings, are produced through selective *absorption*. Selective reflection plays practically no role, except in the colors of metals. The light from an arc lamp which has been multiply reflected from the surface of a gold or copper object exhibits reddish hues when projected onto a screen, with very little white tinting.

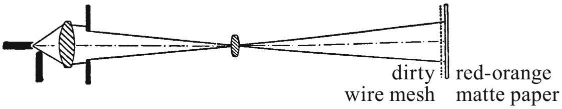


Figure 29.16 The appearance of brilliance or shine (**Video 29.4**)

Video 29.4:
"How brilliance comes about"
http://tiny.cc/zgggoy

29.12 Brilliance and Shine

Often, we see objects as not only colored, but also as shiny. As examples, we could mention polished wood and smooth metal objects, e.g. a copper teakettle. *Shine* or *brilliance can be seen whenever light is scattered in strongly preferred directions*. In this type of scattering, even small motions of the object or the observer change the perceived luminance strongly. The essentials can be shown in a simple demonstration experiment.

In Fig. 29.16, on the right we see a piece of non-shiny, matte red-orange paper. At a few millimeters distance in front of it is a dirty and therefore likewise non-shiny, slightly beaten-up wire sieve. At the left is the light source, an arc lamp. The paper and the sieve can be rotated or swung around together. Every observer believes that he or she is seeing a somewhat dented but still very *shiny copper sheet*. The reason: The shadow of the sieve falls on the paper surface. At certain angles, the observer looks through the mesh of the sieve at the shadows of its wires; the luminance in such directions is small. From a somewhat different angle, he or she looks through the meshes at the unshadowed parts of the paper, so that the luminance is large. 'Shine' is thus *seen*; it is no more a physical property of the object than is its color.

29.13 Shimmering Colors

Shimmering colors are shiny colors in which a small change in the angle of observation or of the incident light causes a strong variation in the hue.

Their hues alone, that is without any directional dependence, can be produced for example with the arrangement shown in Fig. 29.13. We need only replace the template by a screen and use it to mask the main portion of the spectrum, beginning either at its violet or red end.

Shiny, that is strongly directionally-dependent, shimmering colors can be seen for example on the wings of butterflies and beetles, and on some ornamental vases. In these cases, we always find substances which form *layered structures*. In them, either a selective *reflection* of light occurs on a series of equidistant planes (Vol. 1, Fig. 12.41c), or else a selective *transmission* of light occurs; the latter is found even at the boundary between two materials, when either the short-wave portion or the long-wave portion of the natural light is totally reflected. Occasionally, we are dealing with interference filters, like those that were treated in Sect. 20.13.

"This chapter is intended only to stimulate the reader to engage in self study of the very diverse and delightful subject of *color*".

To conclude, we emphasize once more the preliminary remarks in Sect. 29.1: We have not attempted or claimed to give a complete treatment here. This chapter is intended only to stimulate the reader to engage in self study of the very diverse and delightful subject of *color*.

Addendum to Comment C29.4:
Concerning the formation of the color brown (the color which was symbolic of the National Socialists in Germany), we add the following quotes which were collected by Mr. E. Sieker:
1. When the first (brown) SA uniforms appeared in his lecture room, Pohl was not pleased: "Either you participate in the SA, or you study physics – you won't have time for both!" (Source: "*Praxis der Naturwissenschaften Physik*", 6/77, June 15th, 1977, page 159).
2. When Pohl was demonstrating his well-known experiment on the movement of colored ions in KCl crystals, he commented on the side, "As you can see, the brown masses are shuffling along". (This experiment is described e.g. in the 21st edition of Pohl's *Elektrizitätslehre* (1975), in Chap. 25, Sect. 22. Source: *Praxis der Naturwissenschaften Physik* 6/77, June 15, 1977, page 159).
3. The local leadership of the NSDAP (National Socialist or Nazi party) suspected Pohl of being a sympathizer following the assassination attempt on Hitler on the 20th of July, 1944. As became clear only after the War, the NSDAP had not the slightest inkling of Pohl's contacts to the Goerdeler Group of civil resistance – if they had, it would probably have cost him his life – Instead, their mistrust of Pohl, according to statements by the former Rector of the University of Göttingen, was due only to his critical-ambiguous comments on the color brown; for example, that "brown is not a color at all". (Source: H. Becker, H.-J. Dahms, C. Wegeler, "*Die Universität Göttingen unter dem Nationalsozialismus*", 2nd edition, 1998, page 572).
4. During the War Winter of 1944/45, Pohl called indirectly for the capitulation of Germany while demonstrating an experiment. He attached a small white flag to mark a movable object in the experiment, and noted that in 1919, he had been criticized for using a black-white-red flag (the colors of the Old Regime); and later, for using a black-red-gold flag (the colors of the Weimar Republic). Now, he would simply use a white flag. (Source: H. Becker, H.-J. Dahms, C. Wegeler, "*Die Universität Göttingen unter dem Nationalsozialismus*", 2nd edition, 1998, page 572f).

Color Plates

Plate 1

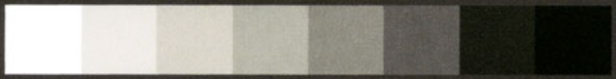


Figure A1. The grey scale

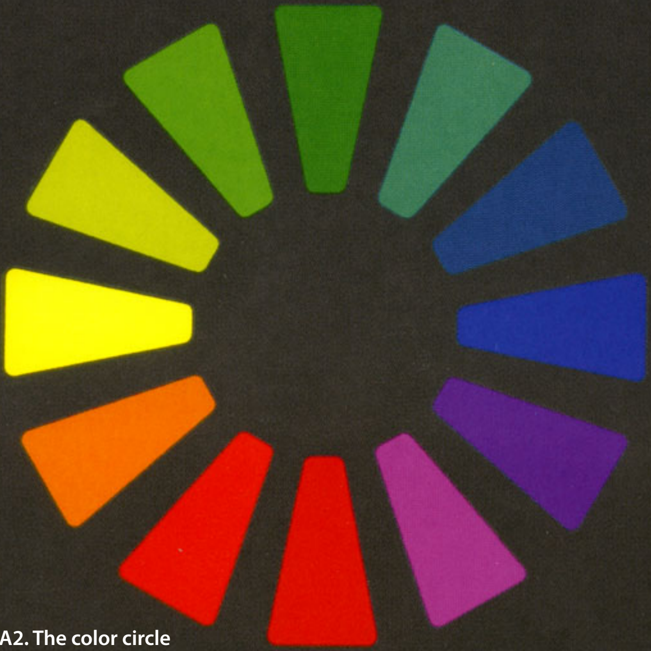


Figure A2. The color circle

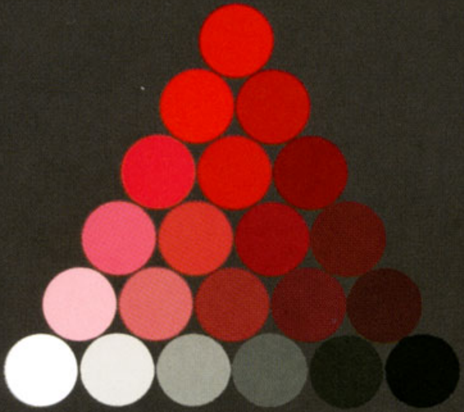


Figure A3. Hering's triangle (E. Hering, 1874)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-50269-4_) contains supplementary material, which is available to authorized users.

Table of Physical Constants

Some important physical constants	(CODATA values from Dec. 2014)	
Avogadro constant	N_A,	$= 6.022140857 \cdot 10^{23}\ \mathrm{mol}^{-1}$
Gravitational constant	G	$= 6.667_4 \cdot 10^{-11}\ \mathrm{N\,m^2/kg^2}$
Electric field constant	ε_0	$= 8.854188 \cdot 10^{-12}\ \mathrm{A\,s/V\,m}$
Magnetic field constant	μ_0	$= 12.566370 \cdot 10^{-7}\ \mathrm{V\,s/A\,m}$
Velocity of light in vacuum	c	$= (\varepsilon_0\mu_0)^{-1/2} = 2.997925 \cdot 10^{8}\ \mathrm{m/s}$
Wave resistance of vacuum	Z_el	$= (\mu_0/\varepsilon_0)^{1/2} = 376.73\ \mathrm{Ohm}$
Relative atomic mass of the proton	$(A)_p$	$= 1.007277\,u^*$
Relative atomic mass of the neutron	$(A)_n$	$= 1.008665\,u^*$
Rest mass of the proton	m_p	$= 1.672621 \cdot 10^{-27}\ \mathrm{kg}$
Rest energy of the proton	$(W_p)_0$	$= 9.382720 \cdot 10^{8}\ \mathrm{eV}$
Rest mass of the electron	m_0	$= 9.101383 \cdot 10^{-31}\ \mathrm{kg}$
Rest energy of the electron	$(W_e)_0$	$= 5.10999 \cdot 10^{5}\ \mathrm{eV}$
Ratio of proton mass/electron mass	m_p/m_0	$= 1836.152$
Elementary electric charge	e	$= 1.602177 \cdot 10^{-19}\ \mathrm{A\,s}$
Specific charge of the electron	e/m_0	$= 1.760366 \cdot 10^{11}\ \mathrm{A\,s/kg}$
Faraday constant	F	$= 96\,485.332_9\ \mathrm{A\,s/mol}$
Boltzmann's constant	k	$= 1.380648 \cdot 10^{-23}\ \mathrm{W\,s/K}$ $= 8.617325 \cdot 10^{-5}\ \mathrm{eV/K}$
Stefan-Boltzmann constant	σ	$= 5.670367 \cdot 10^{-8}\ \mathrm{W\,m^{-2}\,K^{-4}}$
Planck's constant	h	$= 6.626070 \cdot 10^{-34}\ \mathrm{W\,s^2}$ $= 4.135667 \cdot 10^{-15}\ \mathrm{eV\,s}$
Bohr magneton	μ_B	$= he/4\pi m_0$ $= 9.274010 \cdot 10^{-24}\ \mathrm{A\,m^2}\ (= \mathrm{J/T})$
Classical electron radius	r_el	$= \mu_0 e^2/4\pi m_0 = 2.820\,419 \cdot 10^{-15}\ \mathrm{m}$
Rydberg frequency	R_y	$= e^4 m_0/8\varepsilon_0^2 h^3 = 3.289842 \cdot 10^{15}\ \mathrm{s}^{-1}$
Rydberg constant	$R_y{}^*$	$= e^4 m_0/8\varepsilon_0^2 h^3 c = 10\,973\,731.6\ \mathrm{m}^{-1}$
Compton wavelength	λ_C	$= h/m_0 c = 2.426310 \cdot 10^{-12}\ \mathrm{m}$
Sommerfeld's fine structure constant	α	$= e^2/2\varepsilon_0 hc = 1/137.036$
	α	$= \dfrac{\text{Electron's velocity } v_0 \text{ in the smallest H orbit}}{\text{Velocity of light } c}$

[*] u is the atomic mass unit, $u = 1.660539 \cdot 10^{-27}$ kg.
See http://physics.nist.gov/cuu/Constants/index.html.

K. Lüders, R.O. Pohl (Eds.), *Pohl's Introduction to Physics*, https://doi.org/10.1007/978-3-319-50269-4

Solutions to the Exercises

I. Electromagnetism

1.1. $I = 4.98\,\mathrm{A}$

1.2. 745 hours

1.3. The voltmeter is connected in series with a resistance of $99\,\Omega$.

1.4. $R_{\mathrm{shunt}} = 5.025\,\mathrm{m\Omega}$

1.5. $R = U/(I - U/R_{\mathrm{V}})$ (If OHM's law does not hold for the conductor, then R depends on I.)

1.6. Two values of R_2 fulfill this condition: $R_2 = 1\,\Omega$ and $4\,\Omega$!

1.7. The battery has an internal resistance R_{i} of $1\,\Omega$, $U_{\mathrm{I}} = 1.25\,\mathrm{V}$

1.8. $N = 147$

1.9. $I = 0.167\,\mathrm{A}$

1.10. $U = 6.8\,\mathrm{kV}$

1.11. a) $I = 1.25\,\mathrm{A}$, b) $\dot{Q}_2/\dot{Q}_1 = 3$

2.1. a) $\varrho = -\sigma/h$; b) $E = -(1/\varepsilon_0)\varrho(h - x)$, up to the height h, E is negative, and for $x \geq h$, $E = 0$; c) $U_{\mathrm{x}} = -(1/\varepsilon_0)\varrho(x^2/2 - hx)$; d) $U = (1/\varepsilon_0)\varrho h^2/2$

2.2. $C = 356\,\mu\mathrm{F}$

2.3. $C = 0.9995\,\mathrm{nF}$

2.4. $C_{\mathrm{V}} = 0.9\,\mathrm{nF}$

2.5. $N_{\mathrm{e}} = 5.54 \cdot 10^{10}\,\mathrm{cm}^{-2}$

2.6. $C/l = 93\,\mathrm{pF/m}$

2.7. a) 1.37 m from Q_1, on the side away from Q_2; b) 25 cm away from Q_1, between the two charges

2.8. $\int U\,\mathrm{d}t = 10^{-5}\,\mathrm{V\,s}$

2.9. a) $t_{1/2} = RC \ln 2$; b) $C = 44.8\,\mathrm{pF}$

2.10. With $I = \mathrm{d}Q/\mathrm{d}t$ and $U = -Q/C$, we find $\mathrm{d}Q/\mathrm{d}t = -Q/RC + U_\mathrm{g}/R$, $(U_\mathrm{g} = 157\,\mathrm{V})$, with the solution $Q = Ae^{-t/RC} + U_\mathrm{g}C$, $(A = U_\mathrm{max} - U_\mathrm{g},\ U_\mathrm{max} = 220\,\mathrm{V})$. It then follows that $\tau = 1.16\,\mathrm{s}$.

2.11. $W_\mathrm{Batt} = U_0 \int I\,\mathrm{d}t = U_0Q = U_0^2C$, $W_\mathrm{C} = (1/2)U_0^2C$ (Eq. (3.19)). Half of the energy supplied by the battery is therefore converted to heat in the resistor, independently of its resistance! This result however does not hold for $R = 0$, since then the condenser would be charged within an infinitely short time without allowing any energy to be converted to heat. See Chap. 10, “Electrical Oscillations”.

2.12. 17 plates

2.13. $\varepsilon = 2.8$

2.14. 6.5 % of the volume between the plates is filled with metal. This reduces the spacing of the plates effectively by 6.5 %, i.e. the capacitance is increased by about 7 %. This leads to $\varepsilon = 1.07$.

2.15. Corresponding to Eqns. (2.11) and (2.25), the capacitance of the completely filled condenser would be $C = \varepsilon\varepsilon_0(a\,b/l)$, and thus the energy stored in it would be $W_\mathrm{e} = (1/2)\varepsilon\varepsilon_0(U/l)^2 a\,bl$. In the present experiment, the energy thus increases when h increases, by an amount $\Delta W_\mathrm{e} = (1/2)(\varepsilon - 1)\varepsilon_0(U/l)^2 hbl = Fh$. The liquid is thus pulled up into the condenser with a constant force F. h is then found by setting F equal to the opposite force of the weight of the liquid which has been pulled up: $h = (1/2)(\varepsilon - 1)\varepsilon_0(U/l)^2/(\varrho g)$.

2.16. The electrical energy is $W_\mathrm{e} = (1/2)\varepsilon\varepsilon_0(U/l)^2 A\,l$. This leads (compare Eqns. (3.17) and (3.11) in Sect. 3.7) to the force $F = (1/2)\varepsilon\varepsilon_0(U/l)^2 A$.

3.1. The weight of the brass plate is $F = 1.6\,\mathrm{N}$. Using Eq. (3.12), this leads to $l = 50\,\mu\mathrm{m}$ (this corresponds roughly to the thickness of a human hair; see Vol. 1, Sect. 1.2). The roughness of the polished stone causes an average spacing which is however certainly much smaller, so that a considerable fraction of the potential drop must occur within the stone!

3.2. $R = 80\,\mathrm{nm}$

3.3. The capacitance C and thus also the energy W_e vary by a factor n. If $n < 1$, then energy will be pumped back into the current source.

5.1. $\int U\,\mathrm{d}t = 3.2 \cdot 10^{-4}\,\mathrm{V\,s}$, in good agreement with the observed 10 scale divisions

5.2. a) $u = 2.04\,\mathrm{m/s}$; b) $\int U\,\mathrm{d}t = 3.9 \cdot 10^{-5}\,\mathrm{V\,s}$, in good agreement with the observed 1.2 scale divisions $(3.8 \cdot 10^{-5}\,\mathrm{V\,s})$

5.3. a) $B_\mathrm{h} = 2.5 \cdot 10^{-5}\,\mathrm{V\,s/m^2}$; b) $\varphi = 66°$ (In Göttingen, the horizontal component of the earth's magnetic field points to the north and its vertical component points downwards).

5.4. $\nu = 1.06\,\mathrm{Hz}$

5.5. The voltmeter indicates zero! (See also the first paragraph in small print in Sect. 5.5). The pilot could indeed detect an electric field, e.g. by carrying out the experiment of W. WIEN described at the end of Sect. 7.3, but he cannot detect the voltage of $U = 0.8\,\mathrm{V}$ between the wing tips as given by Eq. (5.9), since it will be compensated by the charges which collect on the wing tips. (The "insulated wire" corresponds to the conductor CA in Fig. 7.2).

5.6. The angle between the field vector $\boldsymbol{B}$ and the surface vector $\boldsymbol{A}$ is $\alpha = 2\pi\nu t$. From this, it follows from Eq. (5.10) that $U = -\mathrm{d}/\mathrm{d}t(BAN\cos(2\pi\nu t)) = 2\pi\nu BAN\sin(2\pi\nu t)$, i.e. an AC voltage, as described in Chap. 8 (Fig. 9.3). Note that this result does not depend on the frame of reference.

6.1. a) $I = 93\,\mathrm{A}$ ($\int H\,\mathrm{d}s = 1.85$ scale divisions); b) measured: $H = 1360\,\mathrm{A/m}$, calculated with Eq. (4.1): $H = 1610\,\mathrm{A/m}$; c) $\int H\,\mathrm{d}s = 500\,\mathrm{A}$

6.2. a) Since the magnetic potentiometer measured outside all the remaining coils along a closed path (Fig. 6.7), we find $U_\mathrm{mag} = 0$; b) From a), it follows that $U_\mathrm{mag,i} = -U_\mathrm{mag,a}$.

6.3. If we combine the two methods of measuring the magnetic potentials $U_\mathrm{mag,a}$ and $U_\mathrm{mag,i}$, a closed path results. Around this path, the magnetic potential is always equal to zero unless the path encloses a current. This is the case here, since, as mentioned in Sect. 6.3, each tunnel bored through the material does not pass through individual molecules but rather between them; i. e. no molecular currents can be included. We thus have $U_\mathrm{mag,i} = -U_\mathrm{mag,a}$. The same result follows also from the MAXWELL equation (14.12). (See also Exercise 6.2.)

8.1. $B = 3.37 \cdot 10^{-4}\,\mathrm{T}$

8.2. $M_\mathrm{mech} = 0.1\,\mathrm{N\,m}$

8.3. From the area which is marked with dashed lines, we obtain the flux $\Phi \approx 1.5 \cdot 10^{-6}\,\mathrm{V\,s}$. The magnetic flux measured in Video 5.1 is $8 \cdot 10^{-6}\,\mathrm{V\,s}$.

8.4. $W_\mathrm{magn} = (1/2)\mu_0(N^2/l)I^2\pi r^2$. W_magn increases when l becomes smaller. In order to prevent this, we require a force of $F = 10^{-2}\,\mathrm{N}$.

10.1. $L = 0.55\,\mathrm{H}$

10.2. $L_\mathrm{N} = N^2 L_1$

10.3. $R = 3.466\,\Omega$

10.4. $I_\mathrm{eff} = 0.796\,\mathrm{A}$

10.5. $\nu = 50\,\mathrm{Hz}$

10.6. $L = 123\,\mathrm{mH}$

10.7. $\nu = 97.5\,\text{Hz}$

10.8. $L = 158\,\text{mH}$

10.9. $LC = 2.53 \cdot 10^{-8}\,\text{s}^2$

10.10. a) $Z_{\text{RL}} = 849\,\Omega$, $\varphi_1 = 87.4°$ (the current in the RL branch "lags behind" the applied voltage); b) $I_{\text{RL},0} = 86\,\text{mA}$, $I_{\text{C},0} = 84.85\,\text{mA}$, $I_0 = 86\,\text{mA} \cdot \sin 2.6° = 3.9\,\text{mA}$, $\varphi_2 = 15.2°$; c) It follows from these values that $Z = 73\,\text{V}/3.9\,\text{mA} = 1872\,\Omega$, and from Eq. (10.33), we find $Z = 1840\,\Omega$, in agreement with the value found in b) (compare this value also with the one that can be read off at the maximum of the curve in Fig. 10.20).

10.11. $\Lambda = 0.139$, in agreement with the value in Fig. 10.20.

10.12. a) The active current is 3.77 mA and the reactive current is 1.0 mA; b) $\dot{W} = (1/2)\,U_0 I_0$ $\cos 15° = 0.138\,\text{W}$, $\dot{W}_{\text{RLC}} = (1/2)(I_{\text{RL},0})^2 R = 0.141\,\text{W}$, and thus within the error limits, the values are the same.

10.13. The current in the primary coil is $I_\text{p} = I_{\text{p},0} \sin \omega t$, with $I_{\text{p},0} = U_{\text{p},0} l_\text{p}/(\mu_0 \omega N_\text{p}^2 A_\text{p})$ (Eq. (10.4)). From this we find, with Eqns. (4.1) and (5.4), the flux density B of the magnetic field from the primary coil, and with it, applying the law of induction (Sect. 5.6), the voltage across the secondary coil, $U_\text{s} = N_\text{s}(-\text{d}B/\text{d}t)A_\text{p}$. This leads finally to Eq. (10.39).

10.14. a) Owing to the phase difference of 90° between the current and the voltage, in the primary circuit we expect $\overline{\dot{W}} = 0$. b) With a finite value of R, a current will flow in the secondary circuit. The thermal energy from JOULE heating in the secondary circuit then must come from the primary circuit. The mechanism for this is that the current in the secondary circuit induces an additional current in the primary circuit, so that the phase difference there is no longer 90°, and a finite active current results.

11.1. The voltage across the condenser is $U_\text{C} = Q/C$. The voltage across the coil is $U_\text{L} = L\,\text{d}I/\text{d}t = L\,\text{d}^2Q/\text{d}t^2$. These voltages are equal and opposite at all times: $Q_\text{C}/C + L\,\text{d}^2Q/\text{d}t^2 = 0$. This differential equation can be solved by the trial function $Q_\text{C} = Q_{\text{C},0} \cos \omega t$. This then yields $\omega = \sqrt{(1/LC)}$. Thus, we find that $U_\text{C} = (Q_0/C) \cos \omega t$ and $I = \text{d}Q/\text{d}t = -U_{\text{C},0}\omega C \sin \omega t$.

11.2. $U_\text{L} = L\,\text{d}I/\text{d}t$ (Eq. (10.16)). Therefore, $U_{\text{L},1} = L_1\,\text{d}I/\text{d}t$ and $U_\text{L} = (U_{\text{L},1} + U_{\text{L},2}) = (L_1 + L_2)\,\text{d}I/\text{d}t$. From this, it follows that $|U_{\text{L},1}/U_\text{L}| = L_1/(L_1 + L_2)$. The coils thus act as a voltage divider. An application is shown by the experiment described in Fig. 11.7. With the correct choice of the inductance of the wire loop relative to the overall inductance of the oscillator circuit, a voltage will be induced which causes the lamp to light up without burning out.

11.3. With the values of the capacitance $C = 3.2\,\text{nF}$ and the inductance $L = 0.048\,\text{mH}$, we find the frequency to be $\nu_0 = 400\,\text{kHz}$.

12.1. The rectifier allows only half of the sinusoidal current to pass. The throttle coils in Fig. 12.21 pass only the low-frequency FOURIER components of this current (for FOURIER analysis, see Vol. 1, Sect. 11.3). These components include a direct-current contribution which is proportional to the amplitude I_0, and it will be indicated by the galvanometer.

12.2. a) Using the trial function $E_x = E_{x,0} \sin \omega(-t + z/c)$, superposition with the incident wave gives $E_{tot} = 2E_{x,0} \sin \omega(z/c) \cos \omega t$, that is a standing wave with a node at $z = 0$, as seen in Fig. 12.28. b) The magnetic component of the incident wave is described by the term $B_y = -B_{y,0} \sin \omega(t + z/c)$, and thus its amplitude points in the negative y direction (right-handed coordinate system, wave propagation in the negative z direction). The reflected magnetic component is given by $B_y = B_{y,0} \sin \omega(-t + z/c)$ (in phase with the electrical component; the sign follows again from the right-handed coordinate system). Superposition of the two magnetic components gives $B_{tot} = 2B_{y,0} \cos \omega(z/c) \sin \omega t$, i.e. again a standing wave, but now it has a maximum (wave crest) at the reflecting surface ($z = 0$), with the amplitude $2B_{y,0}$.

12.3. a) $E = 220\,\text{V/m}$, $B = 1.8 \cdot 10^{-4}\,\text{V s/m}^2$, $S = 31.5\,\text{kW/m}^2$; the POYNTING vector points towards the interior of the spiral, as if the thermal power were deposited in the spiral from outside. b) $b = 1.35\,\text{kW/m}^2$, $E = 713\,\text{V/m}$, $B = 2.38 \cdot 10^{-6}\,\text{V s/m}^2$ (also effective values).

12.4. The electromagnetic wave which is radiated from the outside and is used to excite the small dipoles of the water molecules (Fig. 12.6) has a wavelength of $\lambda = 3\,\text{m}$. With its velocity of propagation, $c = 3 \cdot 10^8\,\text{m/s}$, we find its frequency to be $\nu = 10^8\,\text{Hz}$. The dielectric constant of water at this frequency is $\varepsilon = 81$ (Table 13.2 and Fig. 13.11); thus its index of refraction is $n = 9.0$. It follows from this that $R = 0.64$ (at normal incidence). 36 % of the incident radiation thus penetrates into the water; this is sufficient to excite the small dipoles.

13.1. The energy changes from $(1/2)QU_0$ to $(1/2)QU_m$, where $U_m = U_0/\varepsilon$. The excess energy is taken up by the person who places the plate into the condenser.

13.2. The voltage is $U' = E(l - d) + (E/\varepsilon)d$ with the electric field $E = U/l$. Solving for ε, we find $\varepsilon = d/(d + l(U'/U - 1))$. In this example, the result is $\varepsilon = 5$.

13.3. The voltage between the points 1 and 2 is zero when the voltage drop over the capacitance C_1 is equal to that over the resistance R_1. Since the charges Q on the two condensers are then equal, we have $Q/C_1 = I/R_1$ and $Q/C_x = I/R_2$. From this, it follows that $C_x = C_1R_1/R_2$.

13.4. The electric field E experiences no depolarization (Eq. (13.11)). It follows that $D = \varepsilon D_0$. The displacement density in the rod is thus larger than in the space outside it. The concept of "depolarization" refers to the field E (see also "demagnetization", Sect. 14.6)

13.5. $P = 3((\varepsilon - 1)/(\varepsilon + 2))\varepsilon_0 E_0$ (compare this result with the magnetization of a sphere, Eq. (14.18))

13.6. $p_p = 3.04 \cdot 10^{-30}\,\text{A s m}$, and thus only half a large as measured in the gas phase. The reason for this discrepancy is that in deriving the CLAUSIUS-MOSSOTTI equation within the "cavity", the interactions between neighboring molecules were not taken into account; that is, for water, the hydrogen bonds between the water molecules (see H. Fröhlich, "Theory of dielectrics", Oxford University Press (1949), p. 137).

13.7. In paraelectric materials, ε depends on the temperature, both due to the equation of state of ideal gases and to the LANGEVIN-DEBYE theory. For the electric susceptibility, one finds $\chi_e = \varepsilon - 1 = (1/3\varepsilon_0)p_p^2/(kT)^2 \cdot p$ (to distinguish it from the pressure p, the dipole moment is denoted here by p_p). This explains the difference between the two values of χ_e to better than 1 %.

13.8. a) From Fig. 13.11, we read off the dielectric constant at this frequency, thus obtaining the index of refraction $n = 9.0$. It then follows that $\lambda_W = 1.36\,\text{cm}$. b) $\overline{\dot{W}} = 697\,\text{W}$. c) $S = 11.6\,\text{kW/m}^2 = 8.6\,E_e$ (this is about the same as the value of the solar constant on the planet Mercury!) (For reflection losses, see Exercise 12.4).

14.1. $\chi_m = -1.42 \cdot 10^{-5}$. The permeability μ is thus only slightly greater than 1.0, so that the demagnetization effect can be neglected.

14.2. a) Analogously to Eq. (13.15), we have here $H = H_0 + H_m$. Comparing with Eq. (14.18), we then find $H_m = -M/3 = -((\mu - 1)/(\mu + 2))H_0$. b) From Eq. (14.7), it follows that $B_m = (2/3)\mu_0 M$.

14.3. $H_i = 3\mu_a H_a/(\mu_i + 2\mu_a)$; from this, we find $\mu_i = 1$ and $H_a = H_0/\mu_a$: $H_i = 3H_0/(1 + 2\mu_a)$, and for $\mu_a \gg 1$: $H_i \approx (1.5/\mu_a)H_0$.

14.4. $F_0 = 600\,\text{N}$, $F_d = 17.6\,\text{N}$. Due to the hysteresis of the iron (Fig. 14.7), we must consider these values to be only rough estimates.

14.5. a) and b): Following the rapid change of the current during the first second, both measurements show an exponential variation with a relaxation time of $\tau_r = 10\,\text{s}$. c) Here, also, following an initial rapid change, an exponential variation begins, but with a relaxation time of 110 s. Only near the saturation value does the rate of change become more rapid and again approaches a relaxation time of 10 s. For a qualitative explanation: During the initial rapid variation, of which we can see an inkling even in experiment c), the magnetic domains cannot follow. The coil thus has a small inductance during this variation. Thereafter, the reversals of the domains delay the rate of change of the current in the coil in the same manner as the changes of the current in the coil delay its rate of change due to induction. The coil now has a larger inductance because of the iron core. From experiments a) and b), we obtain $L = \tau_r R = 1300\,\text{H}$. In experiment c), the delay of the current changes in the coil is even greater, since in the first 60 seconds, the magnetic domains (remanent magnetization) rotate through 180°, causing the induced voltage which delays the current change to be increased. This results in a still larger inductance: $L = 14\,000\,\text{H}$. Only near the end of the increase in the current, after the remanent magnetization has been completely overcome, does the rate of change again correspond to the same time constant as in experiments a) and b); L is thus again smaller.

14.6. In Exercise 14.2, it was shown for a magnetized ball that the field H_m in its interior is given by $H_m = -NM$ (N is the demagnetization factor). Starting from Eq. (14.14), it can be shown that the same result holds for every ellipsoid of rotation which is magnetized parallel to its axis of rotational symmetry. For the disk with $l/d = 0.1$, we find $N = 0.863$ (Table 13.4). For a long rod ($l/d = \infty$), $N = 0$, and the field H_m is thus zero. From this, with Eq. (14.7) we find for the field B_m in the disk the value $B_m = \mu_0(M - 0.863M) = 0.137\mu_0 M$, and for the rod, $B_m = \mu_0 M$. These flux densities continue into the space outside the material, since div $B = 0$ (MAXWELL's equation (14.13)).

II. Optics

16.1. From the expressions for the law of refraction on both sides of the prism, $\sin\alpha_1 = n \cdot \sin\beta_1$ and $\sin\alpha_2 = n \cdot \sin\beta_2$, we obtain from the sum and the difference: $\sin\alpha_1 \pm \sin\alpha_2 = n \cdot (\sin\beta_1 \pm \sin\beta_2)$. Rearranging, this gives $\sin((\alpha_1 + \alpha_2)/2) \cdot \cos((\alpha_1 - \alpha_2)/2) = n \cdot \sin((\beta_1 +$

$\beta_2)/2) \cdot \cos((\beta_1 - \beta_2)/2)$ and $\cos((\alpha_1 + \alpha_2)/2) \cdot \sin((\alpha_1 - \alpha_2)/2) = n \cdot \cos((\beta_1 + \beta_2)/2) \cdot \sin((\beta_1 - \beta_2)/2)$; dividing these two equations yields: $\tan((\alpha_1 + \alpha_2)/2)/\tan((\alpha_1 - \alpha_2)/2) = n \cdot \tan((\beta_1 + \beta_2)/2)/\tan((\beta_1 - \beta_2)/2)$. With $\beta_1 + \beta_2 = \gamma$ and $\alpha_1 + \alpha_2 = \delta + \gamma$, Eq. (16.8) follows. b) Inserting the values of $\beta_1 = 19.47°$ and $\gamma/2 = 30°$ into Eq. (16.8), we obtain a quadratic equation for $\tan(\delta/2)$, and thus initially two solutions for δ. To find the correct solution, we consider a sketch corresponding to Fig. 16.13. The result is finally $\delta = 47°$.

16.2. The point at which the incident beam in Fig. 16.24 is reflected by the mirror is called A, and the angle of reflection is α. Then we have $AF = FZ = R/(2\cos\alpha)$. For rays near the axis of the mirror (paraxial rays), α is small and so $\cos\alpha \approx 1$. It follows that $f = R - FZ = R/2$.

16.3. A scale drawing corresponding to Fig. 18.1 for a concave mirror with an object y at a distance a and the associated image y' at a distance b allows us to compare similar triangles and obtain $y/y' = f/(b-f) = a/b$, from which it follows that $1/f = 1/a + 1/b$.

16.4. A parallel shift of light beams, even when they have differing wavelengths, leaves the image in the focal plane of the eye unchanged (see e. g. Fig. 18.24, with parallel light beams between the lens and the eye).

17.1. a) $d \approx 100\,\text{m}$; b) $B \approx 11\,\text{cm}$

17.2. $2\varphi_{\text{min}} = 1.04 \cdot 10^{-7}$, $\alpha = 2.28 \cdot 10^{-7}$

18.1. Considering two pairs of similar triangles in Fig. 18.1, we find $y/y' = a/b = f'/(b-f')$. From this, it follows that $1/a + 1/b = 1/f'$.

20.1. a) $m_{\text{max}} = 133$, $x = 21\,\text{cm}$, b) $x_{10} = 1.19\,\text{m}$

20.2. The observation is carried out at an angle of inclination β_{m}, which can be treated as constant for small changes in the order number. Then from Eq. (20.4), we find $d = (1/3)\lambda/2 = 0.1\,\mu\text{m}$.

21.1. a) No change, since over the whole slit, the same path differences are found; b) likewise no change, since the path difference due to the glass, $\Delta = d(n-1)$, is an integral multiple of the wavelength ($\Delta = 145.8\,\lambda$, thus nearly the "order-one position"); c) $\alpha = 8.2°$. since in general the quantities d, n and λ are not known with sufficient accuracy, this calculation shows how one can obtain the two positions in practice.

21.2. Due to the phase difference of $\lambda/2$ between the two halves of the slit, in making the construction, we must rotate the arrows 7 to 12 by 180°; we then obtain the diffraction pattern of the order-two position, that is with extrema at $\sin\alpha = N\lambda/B$, where $N = 1, 3, \ldots$ for the maxima and $N = 0, 2, 4, \ldots$ for the minima. (A more precise calculation shows that the maxima at $N = 1$ (Fig. 21.12) are slightly shifted.)

21.3. The angular range of the principal maximum determined by the width B of the slit ranges over $\sin\alpha = \pm\,\lambda/B$. The maxima which occur through interference with immediately neighboring slits are found at $\sin\alpha_{\text{m}} = \pm\,m\lambda/D$ ($m = 0, 1, 2, \ldots$). Then for a), we observe only three maxima, but for b) there are 39 maxima.

21.4. a) In this case, the whole surface of the grating radiates with the same phase; diffraction occurs only at its edges. b) The diffracted waves which emerge from two neighboring pairs of beams and slits produce a diffraction pattern as shown in Fig. 21.12, lower left. The superposition of the waves emerging from all the beams and slits leads to a sharpening of the maxima (Fig. 22.6), but leaves their positions and relative heights unchanged.

21.5. Since in the order-two position (β), the radiant power is zero for the non-deflected beam ($\alpha = 0$), the power of the light which passes through the gaps and the "grating beams" must be the same. Thus, if we can neglect absorption in the grating beams, they must have the same width, i.e. $L/D = 1/2$.

21.6. $c = m\lambda\nu/(m \sin\alpha) = 875\,\mathrm{m/s}$

22.1. a) $\lambda = 600\,\mathrm{nm}$, b) $n = 1.33$

26.1. a) The equation of motion for free damped oscillations is given by $F_\mathrm{R} + F_\mathrm{D} = m\,\mathrm{d}^2l/\mathrm{d}t^2$, with $F_\mathrm{R} = -\alpha\,\mathrm{d}l/\mathrm{d}t$ (frictional force) and $F_\mathrm{D} = -Dl$ (elastic force). That the expression $l = l_0 e^{-\delta t}$, with $\delta = (1/2)(\alpha/m)$, is a solution of this equation of motion can be demonstrated by substituting it into that equation. For the relation between α and Λ, we obtain with this solution $\alpha = 2\delta m = 2\lambda\nu_0 m$. b) The equation of motion is now given by $F_\mathrm{p} + F_\mathrm{R} + F_\mathrm{D} = m\,\mathrm{d}^2l/\mathrm{d}t^2$, or, after inserting the forces and dividing by m: $\mathrm{d}^2l/\mathrm{d}t^2 + 2\Lambda\nu_0\,\mathrm{d}l/\mathrm{d}t + 4\pi^2\nu_0^2 l = (F_0/m)\cos(2\pi\nu t)$. We now insert the complex trial solution $l = l_0 e^{i(2\pi\nu t)}$ together with its first and second time derivatives into the equation of motion, obtaining $4\pi^2[(\nu_0^2 - \nu^2) + i(\Lambda/\pi)\nu_0\nu] = (F_0/l_0 m)e^{i\varphi}$. From this, applying the mathematical relations $a + ib = r(\cos\varphi + i\sin\varphi) = re^{i\varphi}$ (Eq. (25.27)) and $\sin^2\varphi + \cos^2\varphi = 1$, we obtain equations (26.1) and (26.3). (For literature on "forced oscillations", see e.g. The Feynman Lectures, Vol. 1, Chap. 21-5 (http://www.feynmanletures.caltech.edu/), or http://hyperphysics.phy-astr.gsu.edu/hbase/oscdr.html).

27.1. $(1/2)l_{0,\max}^2/l_{0,\max}^2 = (\Lambda/\pi)^2\nu_0^4/[(2\nu_0)^2(\nu_0 - \nu)^2 + (\Lambda/\pi)^2\nu_0^4]$,
$(1/2)(2\nu_0)^2(\nu_0 - \nu)^2 = (1/2)(\Lambda/\pi)^2\nu_0^4$, $2(\nu_0 - \nu) = H = (\Lambda/\pi)\nu_0$. This result holds for every harmonic oscillator, in particular for an electrical resonator.

27.2. Let the deflection be $x = l_0\ \sin\omega t$; taking the time derivative gives: $\dot{x} = \omega l_0 \cos\omega t$. The kinetic energy is thus $W_\mathrm{kin} = (1/2)m\dot{x}^2 = (1/2)m(\omega l_0 \cos\omega t)^2$. With the average value of $\cos^2\omega t = 1/2$ over a whole oscillation, it follows that $\overline{W}_\mathrm{kin} = (1/4)m(\omega l_0)^2$.

Index